EDGE AI

INTELLIGENT APPLICATIONS DELIVERED AT DATA EDGE

Innovation blueprints, business roi

BY UNOVIE 2024

© About Unovie 2024

At Unovie.AI we strive to be your specialized professional AI Innovation partner, delivering unique AI solutions using our innovative private edge technology, the UnoVie platform. We offer engineering expertise and collaborate with partners to bring AI best practices to your organization, by leveraging your internal domain expertise, we create value-driven outcomes within predictable timelines.

We complement your resources and build extremely reliable, safety critical AI systems with explainability and transparency. We adhere to strict data privacy and regulatory AI security requirements. Our core strengths lie in MedTech, BioSciences, Transportation, Telecom.

Unovie
Aspen Lake One (1911)
Suite 125
13785 Research Blvd
Austin, TX 78750
http://www.unovie.ai

Prologue

If 2023 was the year the world discovered generative AI (genAI), 2024 is the year organizations truly began using and deriving business value from this new technology. Organizations are already seeing material benefits from gen AI use, reporting both cost decreases and revenue jumps in the business units deploying the technology.

Artificial Intelligence (AI) use cases for edge computing with critical response times represent a transformative intersection of cutting-edge technologies tailored to deliver instantaneous, reliable solutions across various sectors.

Edge computing, which processes data at or near the source of generation rather than relying on centralized cloud computing, is instrumental in minimizing latency and improving response times. This integration of AI at the edge is notable for its applications in critical scenarios where real-time data processing is essential, including healthcare, industrial automation, transportation, smart cities, telecommunications, and retail.

In healthcare, AI and Internet of Things (AIoT) devices enhance patient outcomes by enabling real-time data analysis and immediate clinical decision-making. AI algorithms assist radiologists in triaging scans, wearables provide personalized health insights, and IoT-enabled remote patient monitoring ensures timely medical interventions.

These technologies not only alleviate the burden on healthcare providers but also facilitate proactive management of chronic conditions, thereby reducing costs and improving patient care

AI USE CASES

FOR REAL-TIME & MISSION CRITICAL DECISION SUPPORT SYSTEMS

Applications in Healthcare

Integration of Artificial Intelligence (AI) and the Internet of Things (IoT) has been transformative for the healthcare sector, offering novel solutions to longstanding challenges. AI algorithms serve as a second set of eyes for radiologists, highlighting suspicious lesions and fractures to prioritize urgent reviews over benign cases. This collaboration improves the detection of critical conditions, enhancing patient outcomes

AI-Assisted Radiology Triage

Radiologists often face burnout from reviewing an overflow of scans. AI platforms like Qure.AI, Zebra Medical Vision, and MaxQ AI, GE Edison have demonstrated significant improvements in identifying critical findings. Qure.AI increased the detection of critical findings in head CTs by 20%, Zebra Medical Vision flagged pneumonia on ten times more chest x-rays than typically identified by radiologists, and MaxQ AI's algorithm improved stroke detection on head scans by 35%

"This synergy between radiologists and AI ensures quicker identification of life-threatening conditions."

Wearables and IoT Devices

Wearables such as smartwatches harness sensor data that can be processed by Generative AI to provide personalized care suggestions, ranging from dietary recommendations to medication modifications synergy between IoT and Generative AI by introducing a system for monitoring high-risk maternal and fetal health (MFH), capturing clinical indicators via IoT sensors and using a deep convolutional generative adversarial network (DCGAN) for outcome classifications.

Remote Patient Monitoring

IoT-enabled devices provide real-time monitoring and analysis of patient vitals, which is crucial for remote care. These devices can collect data, such as blood glucose levels, and communicate with companion devices like insulin pumps for immediate action

"This capability is particularly beneficial in managing chronic conditions and ensuring timely interventions"

Edge Computing in Healthcare

Modern healthcare systems deploy edge computing devices to enhance clinical decision support (CDS). These devices offer low-latency and low-energy solutions, enabling real-time data gathering and analysis. For instance, in operating rooms, IoT-enabled devices can monitor patient vitals during surgery and alert the surgical team to any changes. In assisted-living facilities, sensors can monitor elderly patients and alert staff to emergencies, such as falls

Improved Patient Outcomes

AI and IoT technologies significantly improve patient outcomes and reduce healthcare costs. Real-time drug dosing adjustments, remote monitoring, and in-home care facilitated by these technologies ensure patients receive timely and accurate medical interventions. These solutions also help alleviate the strain on healthcare systems by enabling the remote management of non-critical patient

Industrial Automation

Artificial intelligence (AI) is revolutionizing industrial automation, offering organizations opportunities to unlock value, enhance efficiency, and achieve greater autonomy. This transformation is driven by labor shortages, plants operating at maximum capacity, and increasing demands for product quality, making AI a top priority for operations, IT, and engineering leaders

Predictive Maintenance

Predictive maintenance powered by edge AI is instrumental in minimizing downtime and repair costs across various industries. By leveraging machine learning algorithms and multi-dimensional sensor data, these systems can predict potential malfunctions before they occur, allowing businesses to schedule maintenance activities proactively

Quality Inspection

One significant area where AI is making an impact is in quality inspection. Vision AI systems automate inspection processes, ensuring consistent quality across different shifts, production lines, and plants. This not only enhances product quality but also allows process specialists to focus on exceptional cases, optimizing their expertise

Smart Factories

The concept of smart factories is being realized through advancements in 5G technology and IoT devices powered by edge AI. These factories leverage automation for various operations, from raw material handling to final product packaging. Autonomous vehicles transport materials within premises, robotic arms assemble components, and quality checks are performed using computer vision-enabled cameras

Data Processing in Edge AI

IoT sensors have become ubiquitous in today's industries due to their ability to collect vast amounts of data from various points throughout the supply chain. Coupled with edge AI, these sensors provide valuable insights into optimizing operations and reducing inefficiencies. For instance, if sensors detect low stock levels at one warehouse and excess stock at another, edge AI processes this data locally to initiate a transfer, preventing shortages and reducing operational costs

> *"The integration of AI in industrial automation demonstrates the potential for continuous improvement and innovation, paving the way for more intelligent and autonomous systems”*

Worker Safety

Worker safety is a critical concern in industrial environments where heavy machinery and hazardous materials are commonplace. Using AI-enabled video analytics, manufacturers can identify unsafe conditions in real time and intervene to prevent accidents. Edge computing is crucial for these life-saving decisions as it enables immediate response times

Future Trends

As the industrial sector continues to adopt AI, edge computing solutions that accommodate real-time inferencing and ongoing model accuracy will become increasingly important. Predictive maintenance, smart factory automation, and enhanced worker safety are just a few examples of how edge AI is transforming industrial automation, promoting efficiency, productivity, and safety across various industries

Transportation

Edge computing has several critical use cases within the transportation sector, especially concerning autonomous vehicles (AVs) and vehicle monitoring systems. These applications leverage cutting-edge technologies, including blockchain, to enhance safety, efficiency, and overall performance.

Vehicle Monitoring

Vehicle monitoring systems incorporate global satellite positioning technology and wireless communication technology to provide comprehensive oversight of vehicle operations. By integrating these technologies with blockchain, various value-added services such as command dispatch, target tracking, emergency alarm, and information release are combined into a non-temperable system. This system allows for effective monitoring of routes, fatigue driving, overloading, and emergency situations, thereby improving the credibility and reliability of the information collected

Accident Management

Accident management in autonomous vehicles focuses on automatic positioning and emergency assistance. Using an onboard computer, wireless communication technology, and global satellite positioning technology, help signals can be sent immediately to rescue organizations in case of an accident. Blockchain can provide relevant information, determine the exact location of the vehicle, and assess the severity of the accident, which is crucial for timely rescue efforts

Risk Assessment

For AVs to operate safely, they must assess and predict the paths of other entities on the road. Traditional risk assessment approaches that focus on trajectory prediction and collision detection are often computationally exhaustive and time-consuming. A more efficient approach involves calculating trajectories and predicting collisions only if a dangerous maneuver or adverse traffic condition is detected. This method treats the autonomous vehicle as part of a broader traffic system, which enhances overall road safety

Decision Making

Decision-making in AVs involves path planning, maneuvering through traffic, and automated parking, among other tasks. The decision unit of an AV predicts the actions of nearby vehicles using stochastic models and probability distributions. This information helps the AV decide on the next course of action based on the predicted probabilities. AI-capable systems, such as speech recognition, steering control, and eye tracking, further enhance the efficiency of the decision-making process. However, the lack of a robust mathematical framework to define reliability and ultra-low latency requirements remains an issue

Mobility Management

Efficient and secure mobility management is crucial for vehicle-to-infrastructure (V2I) communications, especially due to frequent handovers and large-scale vehicular machine-to-machine (M2M) communications. Integrating IPv6 with existing security standards for intelligent transportation systems (ITS) faces several gaps. Research has proposed a vehicular communication architecture that secures the IPv6 Network Mobility (NEMO) using Internet Protocol Version Security (IPVsec) and Internet Key Exchange Version 2 (IKEv2), analyzing performance based on various factors like bandwidth, traffic type, and movement speed

Self-Driving Technology and Telematics

Data from vehicle sensors are stored on the blockchain, enabling all parties to monitor and share safety information more securely and transparently. This reduces the risk of data theft and increases the reliability of vehicle operations.

Smart Cities

With the explosive growth of urban population and the trend of urbanization, the concept of smart cities has been proposed and attracted widespread attention. Smart cities employ intelligent means to reduce energy consumption, enhance energy efficiency, alleviate traffic congestion, ensure urban and resident safety, and improve the overall quality of life. In these environments, numerous hardware devices generate data continuously, including light smart devices for daily life such as smartphones and wearable technology, as well as surveillance cameras and various environmental sensors dedicated to urban security

Infrastructure Management

Embedded within urban infrastructure like traffic lights and surveillance cameras, edge devices analyze data locally to identify traffic congestion, monitor environmental conditions, and track suspicious objects for enhanced security

> *“Localized processing enables cities to improve public services, enhance safety measures, and allocate resources effectively, ultimately fostering a more sustainable and livable urban environment"*

Traffic Optimization

The deployment of sensors and AI in traffic management systems helps optimize traffic flow by monitoring real-time data on congestion and road conditions

> *Double deep Q-networks can be used to discover optimal offloading strategies without prior knowledge, thereby reducing energy consumption and latency of edge computing*

Public Safety

Ensuring the safety of urban residents is another critical aspect. AI and edge computing technologies are employed for activities like violence and crime detection, and urban monitoring, including person re-identification

Energy Management

The trend of urbanization is also prompting a rapid increase in energy consumption, presenting challenges for urban energy management. To address this, AI algorithms, when combined with edge computing, offer superior performance compared to traditional methods. Edge computing allows for faster and more accurate energy consumption predictions and energy management by processing real-time data collected from various urban sensors

Telecommunications

5G is combined with artificial intelligence and augmented reality, the result can be a powerful one-two punch for innovation, both technical- and business-focused. When designing systems or architectures for offices, factories, and homes of the future, the power of AI over 5G -- or 5G enhanced by AI -- cannot be ignored, industry experts point out.

Implementing 5G Technology in Autonomous Vehicles

The implementation of 5G technology in autonomous vehicles (AVs) involves various communication methods to ensure seamless connectivity. Unicast enables direct communication between a pair of user equipment (UE), while groupcast, or multicast, allows a transmitter UE to send messages to a set of receivers that meet specific criteria, such as being part of a group. Broadcast communication entails a single transmitter UE sending messages that are received by all UEs within its transmission range, who then decode and potentially retransmit the message to other UEs in the area

Main Constituents of 5G for V2X

Localization and mapping are fundamental to the reliability of autonomous driving. The navigation task is simplified when the AV can match environmental features perceived through its sensors with a pre-existing map. This process aids in estimating the vehicle's location and detecting obstacles based on discrepancies between the map and sensor data. For instance, LiDAR technology, used for localization and object detection, relies on particle filters to generate and compare point clouds with known maps to enhance data accuracy. Map-building involves identifying, classifying, and integrating unobserved landmarks into the actual map

Federated learning in 5G aims to preserve privacy while training deep neural network (DNN) models in a distributed manner. Instead of sending data to a centralized server, the raw data remains with the clients, and a shared model is trained on the server by accumulating locally computed updates. Optimization is achieved by averaging distributed gradient updates from each client to refine the global model. Stochastic gradient descent (SGD) is commonly used to update local gradients, which are then sent to a central node for a global model update

Data Management and Mobile Edge Computing (MEC)

The exponential growth of connected devices is expected to dramatically increase data generation in vehicular networks. This surge will pose significant challenges in processing and storing vast amounts of data, leading to network congestion, high latency, and reduced bandwidth. Mobile edge computing (MEC) addresses these issues by bringing mobile applications closer to the edge, allowing functionalities to be executed near end users. MEC can meet the strict latency requirements of 5G vehicular communications, delivering ultra-low latency and high bandwidth efficiency. Network function virtualization (NFV) is used to virtualize MEC services, providing cloud-like facilities at the edge

Content Placement at the Wireless Edge

Efficient management of caching service content at the wireless edge can improve the cache hit ratio, reducing duplicate transmissions and latency. Popular content, determined by the proportion of requests, can be cached at base stations.

Retail

Edge AI is transforming the retail industry by enabling personalized experiences, improving inventory management, and enhancing store operations. By leveraging real-time data, retailers can create more engaging and efficient shopping experiences.

Personalized Customer Experiences

Retailers can use edge AI to generate real-time product recommendations and discounts tailored to individual shoppers. These personalized promotions can be delivered via mobile apps, digital signage, or in-store displays, directing customers toward products they are more likely to purchase, thereby increasing sales and fostering customer loyalty

Real-Time Inventory Management

Managing inventory is a critical aspect of retail operations. Edge AI applications and IoT devices provide businesses with real-time data on stock levels, allowing timely decisions regarding replenishment and avoiding overstocking or understocking issues. This system reduces the need for large-scale data storage as most computations occur at the edge, improving performance and reducing latency compared to traditional cloud-based services

Edge AI also helps in reducing shrinkage. In-store cameras and sensors analyze data at the edge to identify and prevent instances of error, waste, damage, and theft. By alerting store associates when inventories are low, these applications help minimize stockouts and improve inventory management efficiency

Improved Store Operations

Edge AI enhances store operations through predictive maintenance. By analyzing multi-dimensional sensor data from various equipment, these systems can predict potential malfunctions before they occur, allowing proactive maintenance scheduling. This reduces downtime and improves the overall efficiency of store operations

Augmented reality (AR) powered by edge AI provides in-store navigation assistance, helping shoppers quickly locate items within large stores. AR markers interact with an app on the shopper's phone, providing directions to desired products and making the shopping experience more convenient

Automated Checkout Systems

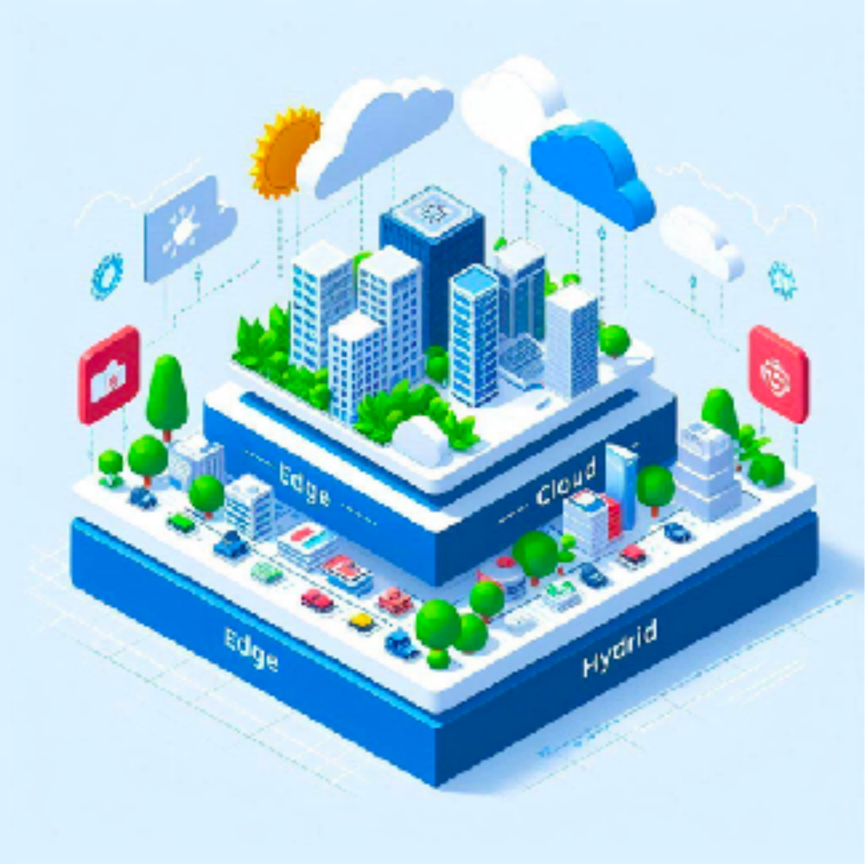
Automated checkout systems powered by edge AI provide seamless shopping experiences, freeing up staff members to focus on other tasks. These systems use sensors and AI to facilitate quick and accurate checkouts, reducing wait times for customers and increasing operational efficiency

Smart Shelves

Smart shelves equipped with edge AI sensors optimize product placement and offer personalized promotions based on real-time analysis of shopper behavior. These shelves monitor inventory levels in real-time, alerting store employees when stock is low or items are misplaced. This reduces the need for manual inventory checks and enhances the shopping experience through targeted advertising

PROS AND CONS

FOR PROCESSING DATA ON THE EDGE USING AI



Historical Background

Corporate IT infrastructure has undergone significant changes over the years. Initially, businesses operated with a static setup where employees worked in large cubicle farms, and data along with business-critical applications were kept close by, often in well-ventilated rooms on the premises or in local data centers

- This model made sense when business operations were confined to single locations, and data had to be quickly and reliably accessible. However, the landscape began to shift as remote work became more common, and businesses expanded into multiple cities and countries

- This growth and the surge in consumer internet usage rendered the traditional on-premises server model less practical. Companies faced challenges scaling their infrastructure, as they had to continually purchase, provision, and deploy new servers to meet the growing demand. Cloud computing services emerged as a solution to these challenges. Services like Microsoft Azure and Amazon Web Services (AWS) allowed businesses to rent server space and scale their operations more efficiently as they grew

- Despite the advantages, these cloud services are centralized, with data centers often located far from end-users. For example, an AWS data center in London is about 330 miles away from Edinburgh, Scotland, and the nearest AWS location for Lagos, Nigeria, is nearly 3,000 miles away in Cape Town, South Africa

- This distance increases latency due to the physical limitations of data traveling through fiber optic cables. The modern solution proposed to address these latency issues is bringing servers closer to the users, echoing the earlier days of on-premises setups but leveraging advanced technologies like edge computing

- This approach aims to reduce latency by processing data closer to where it is generated and consumed.

Comparison with Cloud Computing

Edge computing and cloud computing are two different paradigms for processing data, each with its own set of advantages and disadvantages. While cloud computing processes data in centralized data centers globally, edge computing processes data closer to the source, typically at the edge of the network or within local devices

Data Processing Location

The primary distinction between edge computing and cloud computing lies in the data processing location. Edge computing allows data to be processed closer to the source, which is advantageous for applications requiring minimal latency and quick decision-making

> . In contrast, cloud computing handles data in centralized data centers distributed globally, which can be accessed remotely from anywhere, offering a more centralized and often more scalable approach

Latency

Edge computing offers minimal latency because the data is processed near the point of origin, which is ideal for real-time applications such as autonomous vehicles, gaming, and high-frequency trading

> “The latency in cloud computing is higher due to the physical distance data must travel to reach centralized data centers. Even with high-speed connections, the time needed to send data over long distances can introduce delays, which can be significant for applications that demand immediate responses”

Scalability and Cost

Cloud computing provides high scalability, allowing businesses to scale their computing capabilities up or down based on demand without investing in expensive hardware

”This model also reduces costs associated with maintaining physical infrastructure and offers a variety of pre-packaged services that can support complex analytics and IoT operations”

“Edge computing, while beneficial for reducing latency, may involve higher initial costs for deploying computing resources at multiple locations and can be less scalable compared to cloud solutions”

Accessibility and Global Reach

Cloud computing excels in accessibility and global reach, enabling users to access resources and services from anywhere in the world. This is facilitated by the extensive global network of data centers operated by cloud providers

. Edge computing, while enhancing performance by reducing latency, is more localized and may not provide the same level of global accessibility as cloud computing

Innovation and Agility

Cloud computing offers organizations access to the latest technologies and advanced solutions, promoting innovation and maintaining competitiveness in the market

“This model supports rapid deployment and integration of new services and technologies, which can be crucial for business agility”

“Edge computing, while also offering innovative capabilities, focuses more on optimizing performance for specific applications requiring real-time data processing”

.

Advantages of Processing Data on the Edge

Processing data on the edge offers several significant advantages, making it an increasingly popular choice in various industries.

Lower Operational Costs

Edge processing can substantially decrease operational costs. By aggregating and filtering data at the edge, manufacturers can condense the amount of data that needs to be sent to the cloud by up to 99 percent. This reduction in data transmission leads to lower bandwidth usage and cloud service costs

”Additionally, local data processing utilizes higher bandwidth and storage at lower costs compared to cloud computing, further driving down expenses”

Bandwidth Optimization

Edge computing reduces the need for continuous data transmission to the cloud, thereby conserving bandwidth. Only relevant data is sent to the cloud, optimizing network traffic and enabling more efficient use of network resources

“This is particularly important as the number of connected devices and the volume of data generated continue to grow”

Enhanced Reliability

Edge processing ensures that critical applications remain functional even in the event of network disruptions or limited connectivity to the cloud. Localized processing power provides resilience, which is vital in scenarios where uninterrupted operation is essential, such as remote monitoring and emergency response systems

Improved User Experience

For businesses, lower latency translates into more responsive and capable cloud services and applications. This improvement means reduced employee downtime and almost instant access to the files and information that workers depend on daily. Seamless video conferencing and client presentations without lag are additional benefits, especially when edge computing is combined with 5G technology

Real-Time Decision Making

Edge processing enables real-time decision-making by allowing data to be analyzed and acted upon immediately where it is generated. This capability is essential for applications that require rapid responses, such as fully autonomous vehicles and augmented reality systems

Optimized Data Use

Processing data at the edge allows for quicker and more cost-effective data analysis. By handling data as close to its source as possible, the edge enables actionable insights to be derived swiftly, which can then be used to optimize production and other processes in manufacturing and bey

Disadvantages of Processing Data on the Edge

Processing data on the edge, while offering several advantages, also comes with a set of challenges and disadvantages that organizations need to consider:

Limited Scalability of Individual Devices

Edge computing can scale by adding more devices and nodes, but individual edge devices have inherent limitations in their scalability due to their size, power, and capacity constraints. This can limit the scope of tasks they can handle independently, making it challenging to manage large-scale applications solely on edge devices

Data Consistency and Synchronization

Ensuring data consistency across distributed edge devices and the central cloud or data centers can be challenging, especially in environments where data is constantly being updated. Synchronizing this data without significant latency or bandwidth costs requires sophisticated coordination and data management strategies

Environmental and Physical Risks

Edge devices are often deployed in uncontrolled environments, exposing them to physical damage, theft, or environmental hazards. This necessitates additional physical security and durability considerations, potentially increasing costs and complexity

Resource Constraints

Edge devices and nodes typically have limited processing power, memory, and storage compared to traditional data centers or cloud environments. These constraints can restrict the complexity of the tasks they can perform and the amount of data they can handle, potentially impacting the effectiveness of edge computing solutions in resource-intensive applications

Management Complexity

Deploying and managing a distributed network of edge devices and infrastructure can be more complex than managing centralized data centers. This includes challenges in deploying updates, monitoring performance, and ensuring security across numerous devices and locations, which can increase operational overhead

Managing disparate edge compute, network, and storage systems requires having experienced IT staff available at multiple geographical locations, posing significant financial and logistical challenges

Bandwidth Bottlenecks

Although edge processing reduces the need to transmit large volumes of data to centralized data centers, certain applications still generate massive amounts of data that must be managed efficiently. For instance, radars, sensors, and cameras in autonomous vehicles can generate up to 40 TB of data an hour, necessitating quick and efficient data transfer and analysis

Failure to manage this efficiently can lead to bandwidth bottlenecks, compromising performance and increasing costs

Consistency and Standardization

A lack of standardization and consistency in deploying edge computing solutions can introduce randomness and silos, complicating the management of large-scale distributed infrastructures. Maximizing consistency from the data center out to the edge is crucial but often difficult to achieve

While edge processing offers significant benefits, including reduced latency and enhanced data security, these disadvantages underscore the need for careful planning and management to fully leverage its potential.

Strategies for Effective Edge Computing

Unlocking Data Potential

Edge computing helps you unlock the potential of the vast untapped data created by connected devices. It enables the uncovering of new business opportunities, increasing operational efficiency, and providing faster, more reliable, and consistent experiences for customers. The best edge computing models can accelerate performance by analyzing data locally, which is crucial for real-time application

Local Data Processing

Processing, analyzing, and storing data closer to where it is generated allows for rapid, near real-time analysis and response. As demands from advanced applications grow, centralized data storage in the cloud becomes unsustainable. Edge computing addresses this by leveraging distributed devices, like IoT devices, smart cameras, and industrial PCs, to process data at the source, thus driving exponential growth in data generation and collection

Security Measures

Ensuring security in edge computing is paramount, given the increased attack surface and the decentralized nature of the infrastructure. Effective strategies include preparing for potential security incidents by developing detailed incident response plans and regularly testing and updating these plans. Additionally, educating and training employees on cybersecurity best practices is essential

> *“It's crucial to start with security to avoid the pitfalls of "shiny object syndrome," where excitement over new technology overshadows critical security considerations”*

Automation and Management

With numerous devices deployed at the edge and often limited local IT staff, automation becomes essential. Automated processes can handle mass configuration, respond to events, and update applications centrally. Standardization and consistency, akin to containerization practices, are also critical for managing large-scale distributed infrastructures. Organizations using Kubernetes in data centers are increasingly adopting it at the edge for consistency.

Observability

Observability tools are vital for gaining insights into the performance, health, and operation of edge devices. By collecting, analyzing, and acting upon data generated by edge devices, operations teams can better understand and manage their systems, ensuring optimal performance and quick identification of issues

Distributed Control Plane

The Distributed Control Plane model helps manage large-scale edge data centers by synchronizing control services across various locations. This model can utilize federation techniques or database synchronization to ensure consistent configurations. Collaboration among IT industry groups and standardization bodies is necessary to advance edge computing architectures and address diverse use cases effectively

Future Trends and Developments

Rise of IoT and Real-Time Data Processing

The future of edge computing is closely tied to the proliferation of Internet of Things (IoT) devices and the increasing need for real-time data processing. As more devices become internet-enabled and generate substantial amounts of data, the necessity for edge computing to process this data quickly and efficiently will only grow

“By 2025, the number of connected IoT devices is expected to reach 30.9 billion, producing 73.1 zettabytes of data globally, a 300% increase compared to 2019”

“This surge necessitates robust edge computing solutions for optimal application performance and effective business decision-making”

Hybrid Models

Future developments will likely see the blurring of boundaries between cloud and edge computing, giving rise to hybrid models that leverage the strengths of both

. Such hybrid architectures will integrate the scalability and data processing capabilities of cloud computing with the low-latency, real-time processing advantages of edge computing. These hybrid models will be critical in supporting modern workloads, such as micro-services and machine learning applications, deployed closer to the edge

Edge Computing in Augmented and Virtual Reality

Applications requiring augmented or virtual reality will increasingly rely on edge computing to deliver seamless and immersive experiences

The low-latency access to data provided by edge computing is essential for the real-time processing needs of these technologies, making edge solutions indispensable in these scenarios.

Addressing Infrastructure Challenges

The deployment of modern workloads at the edge presents several challenges that organizations must address to benefit fully from edge computing. Businesses will need to find a perfect balance between their IT infrastructure and end-user needs to harness the potential of edge computing effectively

The strategic alignment of edge devices, such as sensors and IoT gateways, will be crucial for optimal data collection and entry into the network

Continued Growth of Cloud and Edge Computing

The future of computing will witness continued growth in both cloud and edge computing, driven by the rapid uptake of remote working practices, IoT, and AI technologies

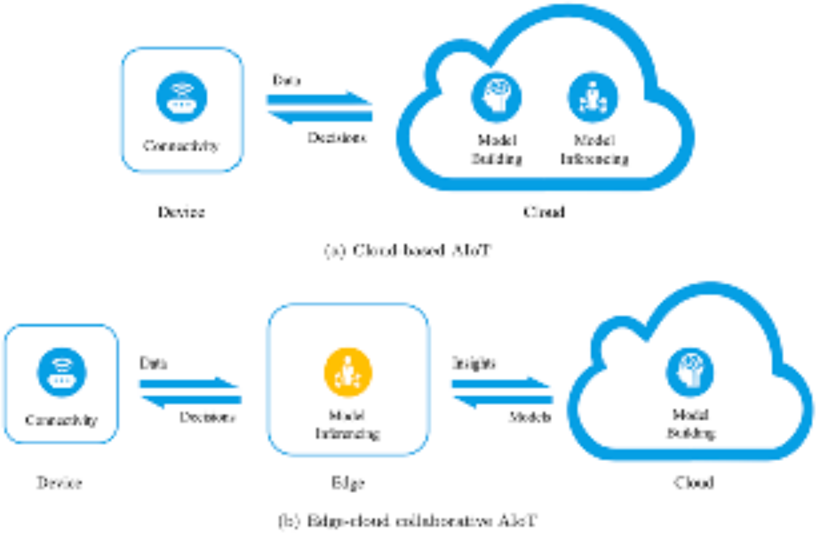
Organizations will increasingly shift to remote working models, leveraging the benefits of "big data" and cloud computing. However, the ultimate trend points towards hybrid models that can seamlessly integrate the two, combining their respective strengths to cater to evolving technological demands.

ARCHITECTURE

ENGINEERING PRINCIPLES OF BUILDING EDGE AI SYSTEMS

Reference Architecture for Edge AIOT

The reference architecture consists of two main stages: Model Building and Inference. The Model Building stage involves collecting data from various IoT devices, processing the data using machine learning algorithms to build models, and storing the trained models. The Inference stage involves deploying the trained models on edge devices or cloud infrastructure to make predictions or decisions based on input data.



(a) Cloud-based AIoT

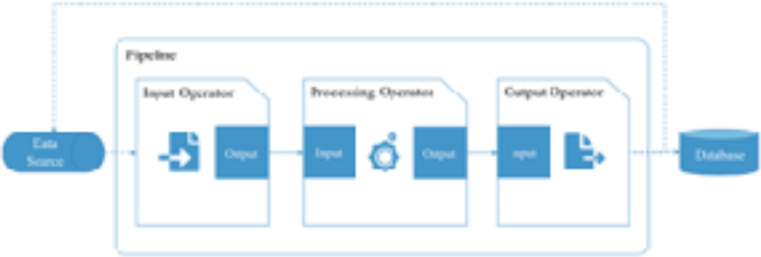
(b) Edge-cloud collaborative AIoT

Model Building Stage

- Data Collection: Collect data from various IoT devices using protocols such as MQTT, CoAP, or HTTP.
- Data Preprocessing: Clean and preprocess the collected data by handling missing values, normalizing the data, and transforming it into a suitable format for model training.
- Model Selection: Choose an appropriate machine learning algorithm based on the problem type (e.g., regression, classification) and data characteristics (e.g., high-dimensional data).
- Model Training: Train the selected model using the preprocessed data. This involves iterating through multiple epochs, adjusting hyperparameters, and evaluating model performance.
- Model Evaluation: Evaluate the trained model's performance using metrics such as accuracy, precision, recall, F1-score, and mean squared error (MSE).
- Model Storage: Store the trained model in a suitable format, such as TensorFlow Lite, Core ML, or ONNX.

Inference Stage

- Model Deployment: Deploy the trained model on edge devices or cloud infrastructure.
- Input Data Processing: Preprocess input data using techniques such as data normalization and feature scaling.
- Model Inference: Use the deployed model to make predictions or decisions based on the input data.
- Output Processing: Post-process the output data, if necessary, by handling outliers or applying thresholding.

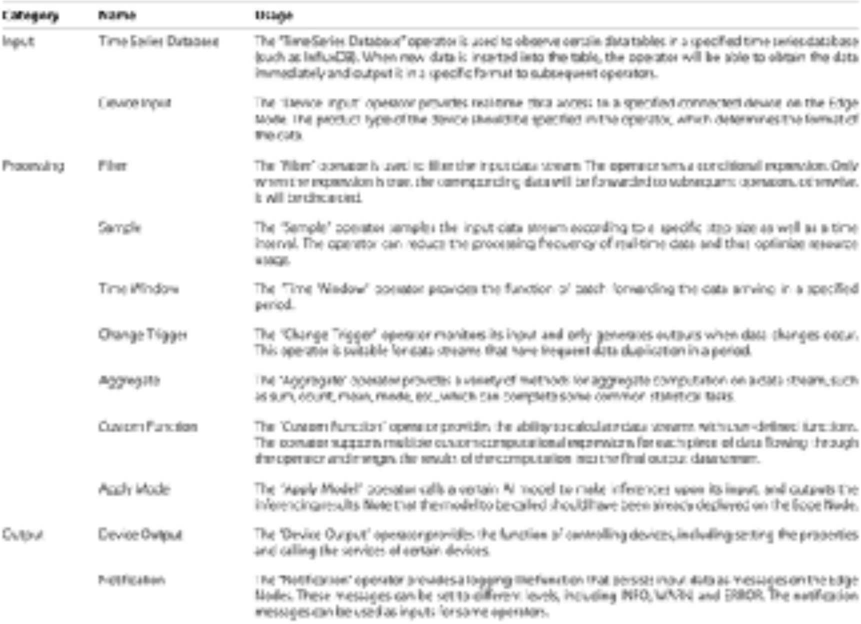


Generalized Pipeline for AIO

Considerations

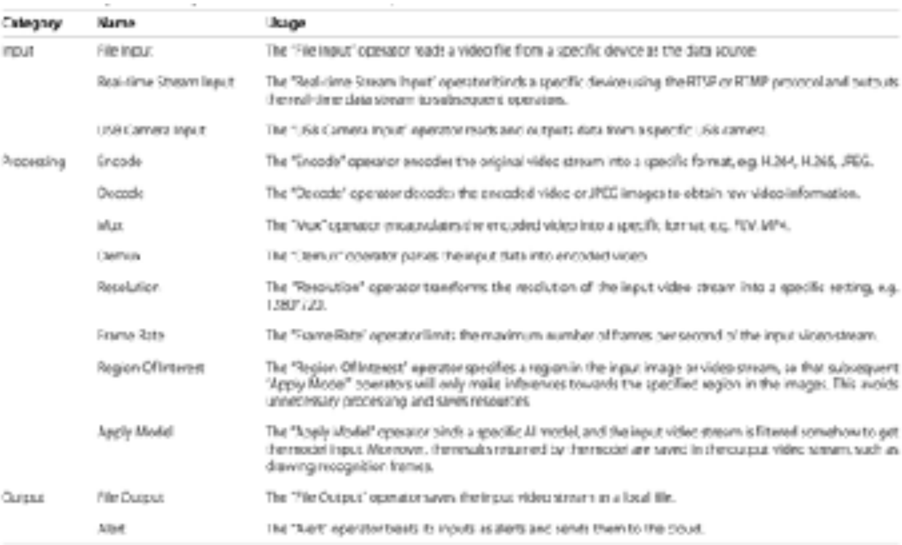
Heterogeneity. The inherent heterogeneity of the devices in a large-scale IoT system makes the connectivity and coordination process very difficult Moreover, due to various protocols, the generated data usually have different formats, sizes, and timestamps, which form a challenging task for further processing, transmission, and storage. Meanwhile, performing general computation on servers with heterogeneous operating systems or runtime also forms a non-ignorable aspect of the challenge.

Accuracy. Algorithms used by AI need to be well developed and tuned to understand and interpret data so that more accurate decisions can be made Additionally, due to the highly dynamic nature of the physical world, a once trained AI model may not perform well all the time, i.e., the models need to be refined using the new-coming IoT data to achieve better performance.

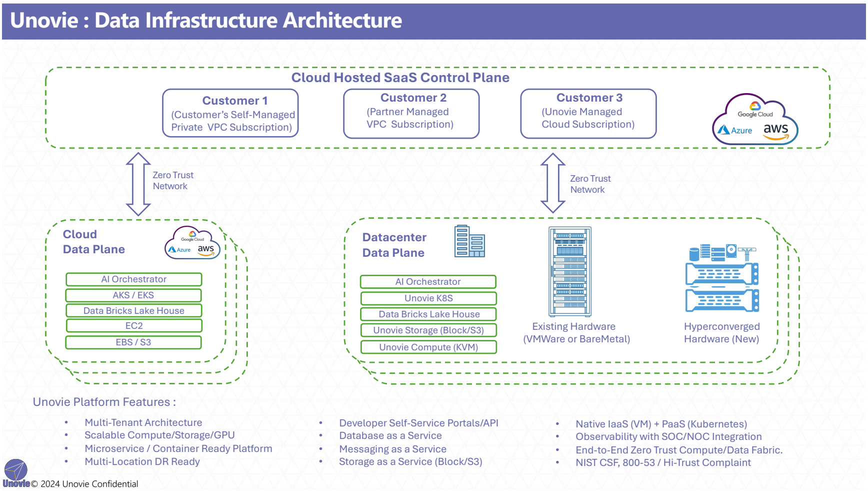


Category	Name	Usage
Input	Time Series Database	The "TimeSeries Database" operator is used to observe certain data tables in a specified time series database (such as InfluxDB). When new data is inserted into the table, the operator will be able to obtain the data immediately and output it in a specific format to subsequent operators.
	Device Input	The "Device Input" operator provides real-time data access to a specified connected device on the Edge Node. The product type of the device should be specified in the operator, which determines the format of the data.
Processing	Filter	The "Filter" operator is used to filter the input data stream. The operator sets a conditional expression. Only when the expression is true, the corresponding data will be forwarded to subsequent operators, otherwise, it will be discarded.
	Sample	The "Sample" operator samples the input data stream according to a specific step size as well as a time interval. The operator can reduce the processing frequency of real-time data and thus optimize resource usage.
	Time Window	The "Time Window" operator provides the function of batch forwarding the data arriving in a specified period.
	Change Trigger	The "Change Trigger" operator monitors its input and only generates outputs when data changes occur. This operator is suitable for data streams that have frequent data duplication in a period.
	Aggregate	The "Aggregate" operator provides a variety of methods for aggregate computation on a data stream, such as sum, count, mean, mode, etc., which can complete some common statistical tasks.
	Custom Function	The "Custom Function" operator provides the ability to calculate data streams with user-defined functions. The operator supports multiple custom computational expressions for each piece of data flowing through the operator and merges the results of the computation into the final output data stream.
	Apply Model	The "Apply Model" operator calls a certain AI model to make inferences upon its input, and outputs the inferencing results. Note that the model to be called should have been already deployed on the Edge Node.
Output	Device Output	The "Device Output" operator provides the function of controlling devices, including setting the properties and calling the services of certain devices.
	Notification	The "Notification" operator provides a logging-like function that persists input data as messages on the Edge Nodes. These messages can be set to different levels, including INFO, WARN and ERROR. The notification messages can be used as inputs for some operators.

Sample Operator for Structured Data



Category	Name	Usage
Input	File Input	The "File Input" operator reads a video file from a specific device as the data source.
	Real-time Stream Input	The "Real-time Stream Input" operator binds a specific device using the RTSP or RTMP protocol and outputs the real-time data stream to subsequent operators.
	USB Camera Input	The "USB Camera Input" operator reads and outputs data from a specific USB camera.
Processing	Encode	The "Encode" operator encodes the original video stream into a specific format, e.g. H.264, H.265, JPEG.
	Decode	The "Decode" operator decodes the encoded video or JPEG images to obtain raw video information.
	Mux	The "Mux" operator encapsulates the encoded video into a specific format, e.g. FLV, MP4.
	Demux	The "Demux" operator parses the input data into encoded video.
	Resolution	The "Resolution" operator transforms the resolution of the input video stream into a specific setting, e.g. 1280*720.
	Frame Rate	The "Frame Rate" operator limits the maximum number of frames per second of the input video stream.
	Region Of Interest	The "Region Of Interest" operator specifies a region in the input image or video stream, so that subsequent "Apply Model" operators will only make inferences towards the specified region in the images. This avoids unnecessary processing and saves resources.
	Apply Model	The "Apply Model" operator binds a specific AI model, and the input video stream is filtered somehow to get the model input. Moreover, the results returned by the model are saved in the output video stream, such as drawing recognition frames.
Output	File Output	The "File Output" operator saves the input video stream as a local file.
	Alert	The "Alert" operator treats its inputs as alerts and sends them to the cloud.



Unovie : Data Infrastructure Architecture

Cloud Hosted SaaS Control Plane

Customer 1 (Customer's Self-Managed Private VPC Subscription)

Customer 2 (Partner Managed VPC Subscription)

Customer 3 (Unovie Managed Cloud Subscription)

Cloud Data Plane

- AI Orchestrator
- AKS / EKS
- Data Bricks Lake House
- EC2
- EBS / S3

Datacenter Data Plane

- AI Orchestrator
- Unovie K8S
- Data Bricks Lake House
- Unovie Storage (Block/S3)
- Unovie Compute (KVM)

Existing Hardware (VMWare or BareMetal)

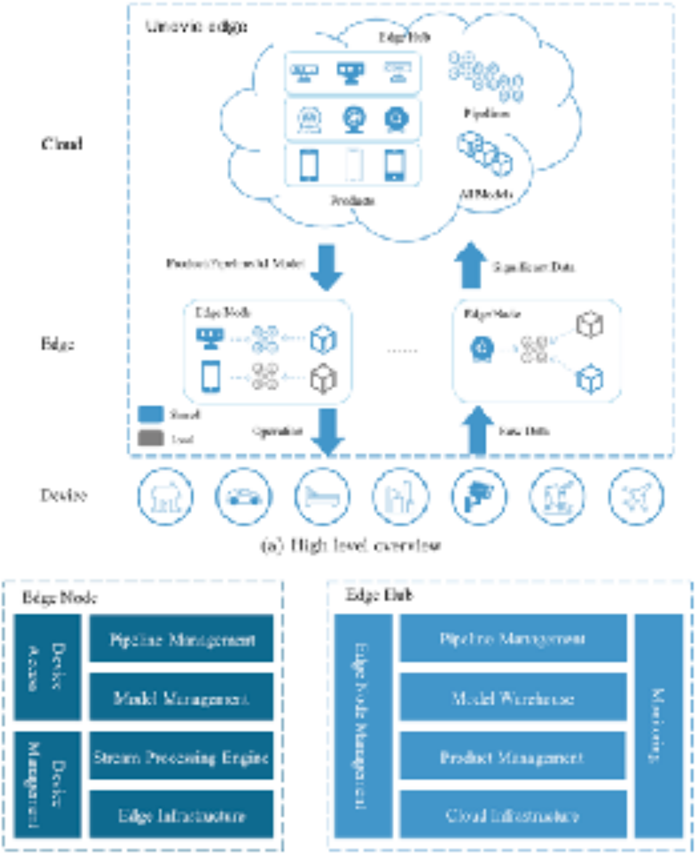
Hyperconverged Hardware (New)

Unovie Platform Features :

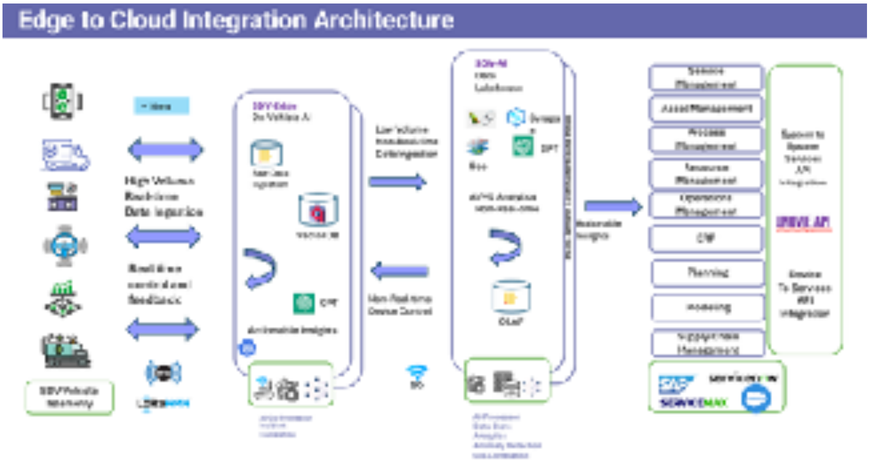
- Multi-Tenant Architecture
- Scalable Compute/Storage/GPU
- Microservice / Container Ready Platform
- Multi-Location DR Ready
- Developer Self-Service Portals/API
- Database as a Service
- Messaging as a Service
- Storage as a Service (Block/S3)
- Native IaaS (VM) + PaaS (Kubernetes)
- Observability with SOC/NOC Integration
- End-to-End Zero Trust Compute/Data Fabric.
- NIST CSF, 800-53 / Hi-Trust Complaint

Unovie © 2024 Unovie Confidential

Sample Operator for Multimedia Data



(a) High level overview


Edge to Cloud Integration Architecture
High Volume Real-time Data Ingestion
Real-time control and feedback
Actionable Insights
GPT
OLAP
Asset Management
Planning
SAP
SERVICEMAX

Platform Architecture

Cloud Orchestration Platform (COP)

The Cloud Orchestration Platform (COP) is a critical component of the Reference Architecture, responsible for managing cloud resources and integrating with edge devices. To support the Edge AI Computing architecture, the COP must meet the following requirements:

Functional Requirements:

Resource Provisioning: Provision and manage cloud resources such as virtual machines (VMs), containers, functions, and storage.

Multi-Tenancy : Support Tenancy Isolation Requirements for SaaS Providers with Multiple Control Planes segregated across multiple business units, departments or locations to reduce the blast radius isolation requirements to support business resiliency

Application Deployment: Deployment of applications on cloud resources, including support for multiple frameworks and languages.

Model Training System: Responsible for training AI models on the processed data using various machine learning frameworks such as Tensor Flow, PyTorch.

Model Serving Platform : Responsible to ensure Versioning of Models, Deployment Descriptors associated with the Device SDK they support and Licensing details

Edge-Cloud Integration: Integrate with edge devices, enabling seamless communication between cloud and edge components.

Data Processing: Data processing and Data Lake-house capabilities, such as streaming, batch, and real-time processing, establish to support Edge AI Computing workloads.

Security and Access Control: Robust security measures, including authentication, authorization, and encryption, to protect sensitive data.

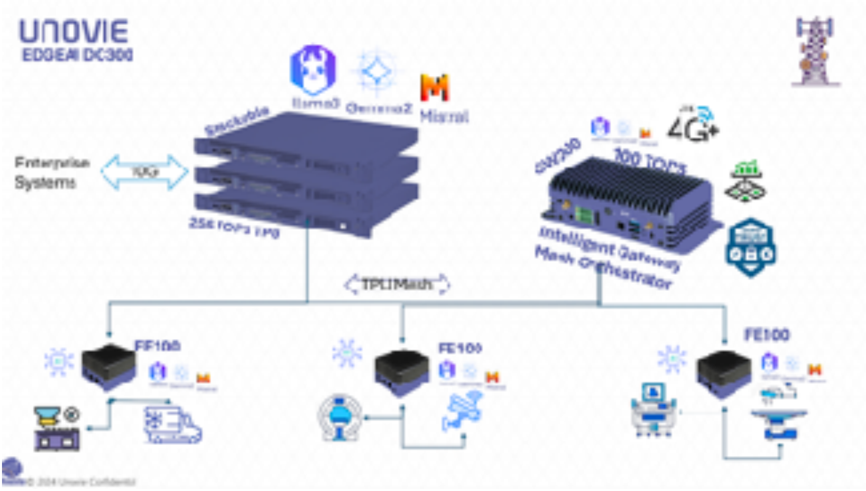
Device Orchestration Platform

The Device Orchestration Platform deployed at Device Hub is responsible for managing edge devices, including firmware updates, device configuration, and monitoring. This platform will interact with the Cloud Orchestration Platform to manage edge applications.

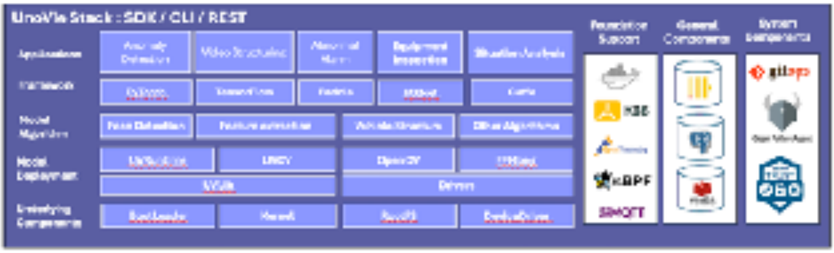
Components:

- **Device Management Agent**: A software component on each edge device that communicates with the Cloud Orchestration Platform.
- **Firmware Update Service**: A service for updating firmware on devices.
- **Device Configuration Service**: A service for configuring device settings, such as network interfaces and security policies.
- **Monitoring Service**: A service for collecting metrics and performance data from devices.
- **Security Service**: A service for managing device security features, such as encryption and access control.

Intelligent AI Mesh Features:



- **Real-time Data Analytics**: Enables real-time insights and decision-making using edge data processing.
- **NPU Scale-out Capacity** : GPU/NPU Offloading capabilities to lease the processing capability from the near by NPU network systems to inference model layers across network in the LLM context.
- **Predictive Maintenance**: Uses machine learning models to predict device failures and schedule maintenance.
- **Energy Efficiency**: Optimizes energy consumption by analyzing energy usage patterns and adjusting device behavior accordingly.
- **Anomaly Detection**: Identifies unusual patterns in device behavior, enabling quick response to potential issues.



Unovie EdgeAI TechStack

AI TechStack

UnoVie Stack, which provides SDK, CLI, and REST interfaces.

Here is a detailed breakdown of the components presented in the stack:

Applications

- **Anomaly Detection:** Applications that identify deviations from normal patterns, often used for security or maintenance.
- **Video Structuring:** Applications that organize video data for efficient storage, retrieval, and analysis.
- **Abnormal Alarm:** Systems designed to trigger alerts based on detected anomalies.
- **Equipment Inspection:** Applications used for assessing the condition of equipment, often using AI for predictive maintenance.
- **Situation Analysis:** Tools for interpreting complex data to understand the current situation, typically in real-time scenarios.

Framework

- **PyTorch:** An open-source machine learning library primarily used for applications such as computer vision and natural language processing.
- **TensorFlow:** An end-to-end open-source platform for machine learning, widely used for both research and production.
- **Paddle:** A deep learning platform developed by Baidu, known for its ease of use in industry applications.

- **MXNet:** A flexible and efficient deep learning framework developed by Apache, suitable for both research and production.
- **Caffe:** A deep learning framework made with expression, speed, and modularity in mind, commonly used for image classification.

Model Algorithm

- **Face Detection:** Algorithms specialized in identifying and locating human faces in images or videos.
- **Feature Extraction:** Techniques used to reduce the amount of resources required to describe a large set of data accurately.
- **Vehicle Structure:** Algorithms focused on analyzing the structure and features of vehicles for various applications.
- **Other Algorithms:** This can include a variety of additional algorithms that may not fit into the other categories but are still relevant to the stack's functionality.

Model Deployment

- **UVRuntime:** A runtime environment designed to execute machine learning models efficiently, possibly specific to the UnoVie Stack.
- **UVCV:** Likely refers to a computer vision library or tool optimized for the UnoVie environment.
- **OpenCV:** An open-source computer vision and machine learning software library, commonly used for real-time computer vision applications.
- **FFMpeg:** A multimedia framework used to decode, encode, transcode, mux, demux, stream, filter, and play almost anything humans and machines have created.

Underlying Components

- **UVLib:** A library component possibly unique to the UnoVie Stack, supporting the functionalities of the stack.
- **Drivers:** Software that allows the stack to interface with hardware components effectively.
- **BootLoader:** A small program that loads the main operating system or runtime environment for the stack.
- **Kernel:** The core part of the operating system, managing system resources and communication between hardware and software.
- **DeviceDriver:** Software that allows higher-level computer programs to interact with a hardware device.

Foundation Support

- **Various Tools and Libraries:**
  - **K3S:** A lightweight Kubernetes distribution used to manage containers in the stack.
  - **OpenTelemetry:** Provides observability into systems, used for logging, metrics, and tracing.
  - **eBPF:** Extended Berkeley Packet Filter, used for running sandboxed programs in the Linux kernel without changing the kernel source code or loading kernel modules.
  - **MQTT:** A lightweight messaging protocol for small sensors and mobile devices, optimized for high-latency or unreliable networks.

General Components

- **Data Storage and Management:**

- **InfluxDB:** A time-series database optimized for fast, high-availability storage and retrieval of time series data.
- **PostgreSQL:** A powerful, open-source object-relational database system with a focus on extensibility and standards compliance.
- **REDIS:** An open-source (BSD licensed), in-memory data structure store, used as a database, cache, and message broker.

System Components

- **GitOps:** A practice of using Git as the single source of truth for declarative infrastructure and applications.
- **Open Policy Agent (OPA):** An open-source policy engine that enables unified, context-aware policy enforcement across the stack.
- **Zero Trust:** A security framework that requires all users, whether in or outside the organization's network, to be authenticated, authorized, and continuously validated for security configuration and posture before being granted access to applications and data.

BUSINESS ROI

FOR PROCESSING DATA ON THE EDGE USING AI

ROI for Edge AI Systems

Return on Investment (ROI) is a crucial metric for evaluating the success of Edge AI systems. Unlike traditional AI projects, which may focus on broad objectives such as improving customer experience or service, Edge AI solutions are often implemented at specific points in the production process, making their use cases more concrete and easier to measure

- This localized approach directly relates to ground-level production, allowing for detailed and particular goals, which in turn make the assessment of value and effectiveness straightforward

- Investing in AI, including Edge AI, should be considered similarly to other business investments like new ERP systems or plants. Each project should have a reasonable payback period and should aim to provide a return on investment; otherwise, it would not be a sensible venture

- Understanding the impact of AI on your business isn't just beneficial; it's essential for grasping how it transforms your revenue streams, reduces costs, and enhances operational efficiency

- Conducting a thorough business impact analysis, which includes examining key performance indicators (KPIs) and other vital metrics, enables well-informed, data-backed decisions that drive the business forward

- The decentralized nature of Edge AI empowers businesses to gain better control over their data and implement robust security measures tailored to specific use cases, ensuring the integrity and confidentiality of their data

- This is increasingly important as the proliferation of IoT devices continues to generate vast amounts of data, necessitating local processing for effective management

- Moreover, Edge AI offers multiple benefits such as improved operational efficiencies, enhanced user experiences, and accelerated business transformation

- These advantages not only contribute to a positive ROI but also support the long-term goals of businesses by improving performance and competitiveness. For example, surveillance cameras and smart video systems that utilize Edge AI for video analytics can process data locally, reducing the need for bandwidth and cloud storage while improving real-time decision-making capabilities

Components of ROI in Edge AI Systems

Understanding the components of ROI (Return on Investment) in Edge AI systems is crucial for businesses aiming to optimize their investments and achieve substantial financial returns. Edge AI, the convergence of edge computing and artificial intelligence, offers unique advantages over traditional AI systems that can significantly impact ROI calculations.

Upfront Costs

The initial investment in Edge AI systems includes several upfront costs. Businesses must consider expenses related to high-speed hardware and specialized software, which can be substantial. For instance, servers alone may cost over $10,000

> “*Additionally, integrating the Edge AI solution into the company's ecosystem often requires developers to customize and align the technology with the brand identity, with integration costs potentially starting at $4,500”*

Training and Support

Ongoing training and support are essential to ensure seamless operation and troubleshooting. These expenses can vary, but companies might expect to invest around $2,000 to provide adequate support as the customer base grows and usage patterns evolve

> "*Furthermore, additional costs might include ongoing maintenance and optimization efforts, typically around $400 per month, to address performance issues and adapt to user feedback*"

Efficiency and Power Consumption

Edge AI systems are designed to be power-efficient, which can significantly reduce operating costs over time. Although the initial cost of integrating a power-efficient architecture might be higher—up to 10% more expensive—the continued benefits of lower operational costs make this investment worthwhile

> "*By minimizing energy consumption, Edge AI solutions contribute to lower overall costs, enhancing ROI*"

Data Security and Integrity

The decentralized nature of Edge AI provides enhanced data security and integrity. Edge AI systems enable businesses to implement robust security measures tailored to specific use cases, using advanced encryption techniques and anomaly detection algorithms to protect data

> "*This improved security reduces the risk of data breaches and associated costs, thereby positively impacting ROI*"

Real-Time Decision Making

Edge AI's ability to process data and make decisions in real-time offers another significant advantage. For example, autonomous vehicles and smart traffic lights can quickly respond to changing conditions, improving efficiency and safety in the Internet of Vehicles (IoV) network

> "*This rapid decision-making capability not only enhances operational efficiency but also reduces potential costs associated with delays and errors”*

Industry-Specific Applications

The cost and ROI of Edge AI systems can vary widely depending on the industry. Healthcare applications, for instance, may range from $20,000 to $50,000, while fintech applications could cost between $50,000 and $150,000 due to differing requirements and regulatory environments

> "*The industry-specific nature of these applications means that ROI calculations must be tailored to the particular needs and benefits within each sector”*

Concrete Use Cases

Edge AI use cases are often more concrete and directly related to ground-level production processes, making it easier to measure the value of the solution. Unlike broader AI goals such as improving customer experience, Edge AI implementations have particular and detailed objectives, allowing for more precise ROI calculations

> "*This specificity ensures that businesses can accurately assess the financial benefits of their investments”*

UNOVIE

AUSTIN TEXAS

HTTP://WWW.UNOVIE.AI

PRIVATE.EDGE.AI

BUILT FOR ISV AND IS PARTNERS