Digital Object Identifier

AI in Edge Computing

SAI VARSHITH GANDU[1], TARUN TALASILA[2], BALAJI CHAPPIDI[3] and SOWGOTO RAHA SUNNY[4]
[1]University of North Texas, Denton TX 76201 USA (e-mail: SaiVarshithGandu@my.unt.edu)
[2]University of North Texas, Denton TX 76201 USA (e-mail: TarunTalasila@my.unt.edu)
[3]University of North Texas, Denton TX 76201 USA (e-mail: BalajiChappidi@my.unt.edu)
[4]University of North Texas, Denton TX 76201 USA (e-mail: SowgotoRahaSunny@my.unt.edu)

ABSTRACT Artificial Intelligence (AI) and Edge Computing converged has been a game changer in the way data could be processed, analyzed and utilized in live environments in real time. With the increasing generation of mobile and Internet of Things (IoT) device data, existing cloud computing models based on traditional computing are experiencing severe latencies, bandwidth limitations and privacy issues. By bringing computational power closer to the data source, it mitigates these challenges the AI driven applications can function with minimal delay. In this paper, we first review the state of the art in AI/edge integration, characterizing how federated learning, deep reinforcement learning, and model compression are driving efficient AI deployment on resource constrained edge devices. In this article we run through the fundamental challenges with edge AI including scaling and energy efficiency, and we dive into some of the burgeoning solutions aimed at overcoming these obstacles. Applications such as healthcare, autonomous vehicles, and smart city infrastructure show the potential for edge AI to reinvent applications by enabling low latency, context aware and energy efficient solutions. The advent of 5G networks, alongside the advance of IoT ecosystem, promises future of AI at the edge full of great opportunity for industry and research, which may open the doors for more intelligent, adaptive and secure AI systems of the future. The main objective of this thesis is to frame the literature on edge AI, offer a roadmap for future research work on realizing edge AI, and identify techniques to fill current limitations for developing edge AI[2].

I. INTRODUCTION

ARTIFICIAL Intelligence (AI) has seen substantial growth this era, due to new and powerful machine learning algorithms and significant computational power available along with large amounts of data. Many sorts of AI have entered numerous domains including individualized recommendations, advanced robotics, computer vision and computer language processing, changing industries and the way life is lived. Central to modern life and applications with real business value include smart assistants, autonomous vehicles, predictive analytics, enhancing productivity and efficiency. To the contrary, due to the proliferation of mobile devices and Internet of Things (IoT) devices, this has led to an unprecedented surge of data created at the network edge. Yet, this trend poses a major challenge to existing traditional centralized cloud computing models, that are typically unable to quickly process the vast amounts of real time data in the presence of latency, bandwidth constraints, and privacy. For instance, in systems that reside in the cloud you will need to transmit your data to a remote server for processing where the delay will be unacceptable for critical applications such as autonomous driving or real time health monitoring.

Decentralizing data processing, through edge computing, has emerged as a significant paradigm to solve these problems. This shortens Latency, conserves Bandwidth, and solves privacy issues by allowing the data to remain on the edge of the networks. Being cloud to edge computing, the shift means we can do real time analysis and decision making to run AIpowered apps locally without always connected to centralized servers. Thus, Edge AI or Edge Intelligence (EI) refers to the combination of both the AI techniques and edge computing to empower smart devices with new capabilities as fast, efficient, context aware services[1].

II. BACKGROUND AND MOTIVATION

In this section, we discuss the evolution of AI and edge computing, examples of why models are limited in the cloud and the motivation to shift AI processing to the network's edge:

A. THE RISE OF EDGE COMPUTING

The increasing number of IoT and mobile devices requires logical processing of the data locally. However, traditional cloud computing models are powerful but can not support leading real time data analysis demands while imposing latency. This is where edge computing steps in by moving computation closer to the data source, thereby significantly lowering response times, saving bandwidth, and also improving on the data privacy. With edge computing, the paradigm is shifted, as the computation is localized to edge nodes such as IoT devices, routers and micro data centers, for faster and more secure data handling

B. ARTIFICIAL INTELLIGENCE IN EVOLUTION

Deep learning, in particular, has seen a tremendous growth with AI in general: machines can now do tasks like image recognition, language parsing, and real time decision making with very high accuracy. But these applications are based on cloud systems, where heavy workloads, such as model training and inference, happen in centralized data centers.

C. CLOUD-BASED AI CHALLENGES

Cloud-based AI solutions have certain limitations that make them unsuitable for real-time, high-bandwidth, and privacy-sensitive applications:

- **Latency:** Models that rely on data travelling over networks to distant servers, often bouncing off multiple hops to arrive at their destination, can have unacceptable delays for real-time applications like autonomous driving and healthcare monitoring.
- **Bandwidth Constraints:** Transmitting enormous amounts of data from IoT devices to centralized servers isn't feasible in high-data environments such as smart cities.
- **Data Privacy:** Sensitive data often needs to be processed locally to ensure privacy and mitigate security vulnerabilities.

D. MOTIVATION FOR AI AT THE EDGE

The integration of AI with edge computing—termed Edge AI—offers several key benefits:

- **Real-time Processing:** AI models can now make decisions locally with near-zero latency, enabling applications like autonomous vehicles and industrial automation.
- **Reduced Bandwidth Usage:** Processing data at the edge reduces the need to transmit large amounts of data to the cloud, conserving bandwidth.
- **Improved Data Privacy:** Local processing avoids sending sensitive data to public cloud servers, reducing the risk of exposure and potential data security vulnerabilities.

E. EDGE AI IS BEING DRIVEN BY 5 KEY TECHNOLOGIES

Many technological solutions have been developed to facilitate the deployment of AI to edge devices, which are usually resource-constrained. These include:

- **Model Compression Techniques:** To run on devices with limited memory and processing power, AI models are made smaller through techniques like quantization, pruning, and knowledge distillation.
- **Federated Learning:** Instead of sharing raw data, federated learning enables collaboration between edge devices to train a shared AI model without directly sharing sensitive data.
- **Edge-specific AI Hardware:** Specialized AI hardware, such as Google Edge TPU and NVIDIA Jetson, has been developed to improve the performance of AI models in edge environments.

F. EXAMPLE APPLICATIONS OF EDGE AI

Edge AI finds practical applications across various domains:

- **Healthcare:** AI can analyze sensor data locally in medical devices, enabling real-time patient monitoring and quick diagnostics, making crucial decisions without waiting for cloud processing.
- **Autonomous Vehicles:** AI models deployed within autonomous cars process sensor data in real-time, eliminating dependence on remote servers and enabling split-second decision-making.
- **Smart Cities:** City management, including traffic control and power distribution, is being optimized by processing enormous amounts of data from traffic cameras, energy grids, and surveillance systems.

III. KEY CHALLENGES AND SOLUTIONS IN EDGE AI

Deploying AI in edge computing environments is inherently challenging due to the resource-constrained nature of edge devices, the need for data privacy, and the dynamic network environment. This section evaluates critical challenges and discusses solutions to address them [5].

A. SCALABILITY AND RESOURCE CONSTRAINTS

Most edge devices, such as IoT sensors, mobile phones, and routers, are computationally constrained, memory scarce, and energy-limited. Unlike centralized cloud servers, these devices are not designed for intensive AI model training or inference.

1) Challenges

- **Limited Computing Power:** Large AI models, such as deep learning models, require substantial computational resources, which edge devices often lack.
- **Memory Limitations:** Edge devices generally have limited storage, making it challenging to host large AI models.
- **Energy Consumption:** Running complex AI models on devices with limited battery life, such as mobile phones,

quickly drains the battery, reducing device longevity and usability.

2) Solutions

- **Model Compression:** Techniques such as quantization, pruning, and knowledge distillation reduce model size with minimal accuracy loss.
- **Edge-specific AI Hardware:** Devices like Google's Edge TPU and NVIDIA's Jetson are designed for efficient AI computation at the edge with lower power consumption.

B. LATENCY AND REAL-TIME PROCESSING

Real-time applications, such as autonomous vehicles and healthcare monitoring, require decisions within milliseconds. Relying on centralized cloud computing introduces high latency, hindering time-critical operations.

1) Challenges

- **Network Delays:** Transmitting data to remote servers for processing introduces delays that are unacceptable for real-time systems.
- **Limited Bandwidth:** Bandwidth constraints in remote or data-intensive environments limit the feasibility of constant communication with cloud servers.

2) Solutions

- **On-device AI Processing:** Processing data locally at the edge minimizes latency and allows for near-instantaneous decision-making.
- **Edge Caching:** Frequently used data and computations can be cached locally, reducing network congestion and improving response times.

C. DATA PRIVACY AND SECURITY

Edge devices often process sensitive data (e.g., healthcare or financial data), making privacy and security paramount. Transmitting raw data to the cloud increases the risk of data breaches.

1) Challenges

- **Data Privacy:** Centralized cloud servers handling sensitive data increase the risk of privacy breaches.
- **Security Threats:** Distributed edge devices are more susceptible to physical tampering and cyberattacks due to limited security protections.

2) Solutions

- **Federated Learning:** This decentralized approach trains global models without sharing raw data, enhancing privacy.
- **Advanced Security Solutions:** Integrated encryption techniques, secure booting, and intrusion detection systems protect edge devices from threats.

D. ENERGY EFFICIENCY

Edge devices, such as IoT devices and mobile phones, are constrained by limited power resources, making energy-efficient solutions essential for sustainable AI deployment.

1) Challenges

- **Energy Consumption:** AI models, especially deep learning networks, are computationally intensive and power-hungry, unsustainable for battery-powered devices.

2) Solutions

- **Efficient AI Models:** Techniques such as pruning and quantization, as well as computation reduction, decrease energy consumption.
- **Adaptive Computation:** Models adjust complexity based on available energy, such as using fewer neurons during low-energy scenarios.

E. DYNAMIC NETWORK CONDITIONS

Edge computing environments are dynamic, with fluctuations in network quality, device availability, and connectivity, posing challenges for consistent performance.

1) Challenges

- **Variable Network Quality:** In rural or remote areas, fluctuating network connectivity affects data reliability and AI model performance.
- **Device Mobility:** Mobile edge devices make stable communication and computation challenging.

2) Solutions

- **Edge Caching and Preprocessing:** Local caching of data and computations minimizes reliance on real-time network conditions.
- **Distributed Learning Models:** Techniques like federated learning and reinforcement learning allow devices to adapt and share computational loads.

IV. ROLE OF WIRELESS NETWORKS IN MAKING EDGE AI POSSIBLE

Wireless networks are enablers of seamless Edge AI system operations. Edge AI generally refers to the deployment of artificial intelligence algorithms and models at the edge of the network, closer to the source of the data, rather than depending on centralized cloud servers [8]. Such an edge deployment enables faster and more efficient processing, helping to keep latency low and improving real-time decision-making. However, for Edge AI to function effectively, reliable high-speed wireless communication is essential.

A. REAL-TIME DATA TRANSMISSION

Running large volumes of data from edge devices into AI models using wireless networks, particularly high-performing ones like 5G with near real-time latency, is be-

coming critical. For applications such as autonomous vehicles, healthcare monitoring, and smart cities, the capability to send data from sensors like cameras, LIDAR, or IoT devices quickly for processing by local units or cloud-based systems is crucial for fast decision-making. This is particularly important in applications that require split-second decisions, such as predicting when industrial machinery will need maintenance or detecting anomalies in health monitoring systems.

B. LOW LATENCY - HIGH THROUGHPUT

One of the major reasons wireless networks, notably 5G, are so highly sought after is due to their ability to provide ultra-low latency with high throughput. Latency minimization is especially critical for Edge AI applications involving robotics or augmented reality. Low-latency wireless networks allow AI algorithms to process information in real-time at the edge, reducing the need for distant data centers. This decreases latency and avoids bandwidth congestion, enabling devices to handle more traffic with increased efficiency without degrading service quality.

C. FLEXIBILITY AND MOBILITY

Wireless networks offer inherent mobility and flexibility compared to wired infrastructure. Edge AI applications in dynamic environments require devices to communicate and process data while moving, such as in autonomous vehicles, drones, or precision farming in agriculture. Wireless networks keep these devices connected, enabling local processing and adaptation to changing conditions, as opposed to relying on a fixed, physical connection. This flexibility makes wireless communication ideal for applications where edge devices are spread over large areas and/or need to be mobile.

D. SCALABILITY

In many Edge AI networks, a high volume of distributed devices, such as sensors and IoT devices, may be involved. Newer wireless networks, like 5G, support a massive number of concurrently connected devices. This scalability is critical for applications like smart cities or Industrial IoT (IIoT), where thousands of devices must communicate with edge servers in real-time. The ability to efficiently process and analyze the data obtained from these devices for decision-making is crucial in such scenarios.

V. WIRELESS NETWORK ARCHITECTURE FOR EDGE AI

The core of Edge AI systems relies on the architectures from wireless networks, where the performance of wireless and low-latency communication between devices and Edge computing resources plays a crucial role. The architecture chosen will directly impact the efficiency, scalability, and reliability of AI applications at the edge of the network. Some of the core wireless network architectures that are capable of enabling Edge AI effectively include 5G, Wi-Fi 6, and LPWAN. Each of these architectures has specific characteristics that make it suitable for different applications in Edge AI.

A. 5G: SUPERSPEED CONNECTIVITY AT LOW LATENCY

5G is revolutionizing wireless communication and is emerging as a key enabler for Edge AI, particularly for applications that require high data throughput, low latency, and massive connectivity. Its advancements make it uniquely suited to support AI systems deployed at the edge, where real-time processing and decision-making are crucial.

Key Features of 5G for Edge AI include:

- **Low Latency:** One of the most critical advantages of 5G is its very low latency, as low as 1 millisecond. This is crucial for applications that need real-time decisions, such as AI-powered autonomous vehicles, real-time industrial automation, and augmented reality (AR). For example, in autonomous driving, milliseconds can make the difference between avoiding a collision or not. The ability to process data from sensors and cameras in real time is paramount.
- **High Throughput:** 5G provides data transfer speeds as high as 10 Gbps, supporting high-bandwidth tasks such as real-time video processing, AI-driven surveillance systems, and edge computing for smart cities. This is especially useful in AI applications that involve transmitting large datasets in short periods, such as remote surgery, where large video streams must be processed instantaneously for a surgeon's interaction with patients over long distances [7].
- **Massive Device Connectivity:** Another key feature of 5G is its ability to support up to 1 million device connections per square kilometer. This becomes essential in large-scale IoT networks, where hundreds of sensors and multiple devices need to work together, such as in smart cities or smart factories. 5G ensures that AI systems at the edge can manage and process data from a vast array of connected devices without performance degradation.

Basic Suitability for Edge AI:

5G is particularly well-suited for Edge AI, as it supports highly complex, real-time AI models with minimal reliance on centralized cloud computing. The combination of low latency, high throughput, and massive device connectivity allows advanced AI applications to operate in real time with minimal delays, even in environments that require frequent updates, such as industrial automation, smart grid monitoring, and smart transportation systems. With 5G, Edge AI applications can process data locally and make split-second decisions without needing to send data to distant servers in the cloud. This not only reduces latency but also eases bandwidth congestion, ensuring faster, more reliable AI-powered systems. Therefore, 5G is a crucial technology for the growth and success of Edge AI across various industries.

B. WI-FI 6: HIGH-EFFICIENCY WIRELESS COMMUNICATION

Wi-Fi 6 (also known as 802.11ax) is the newest generation of Wi-Fi designed to meet the demands of modern networks, particularly in high-density and heavy-traffic environments. It offers significant improvements in speed, efficiency, and capacity over its predecessor, Wi-Fi 5 (802.11ac), making it a key enabler for Edge AI applications in environments such as smart homes, industrial automation, and IoT-heavy settings.

Key Features of Wi-Fi 6 for Edge AI:

- **Higher Throughput:** Wi-Fi 6 supports data speeds up to 9.6 Gbps, compared to 3.5 Gbps with Wi-Fi 5. This higher throughput is crucial for Edge AI applications that require high data rates, such as AI-based video surveillance, real-time analytics, media streaming, and high-resolution imaging. For example, in smart cities, Wi-Fi 6 can handle the continuous streaming of high-resolution videos from numerous AI-powered cameras for real-time analysis, making it essential for such applications.
- **OFDMA (Orthogonal Frequency-Division Multiple Access):** A key innovation of Wi-Fi 6, OFDMA enables multiple devices to transmit data simultaneously on the same channel by dividing the channel into smaller sub-channels. This enhances efficiency, particularly in environments with many connected devices, like smart factories or crowded offices. In Edge AI applications involving numerous IoT devices, OFDMA ensures each device gets sufficient bandwidth without interference, leading to optimal system performance.
- **TWT (Target Wake Time):** TWT is a power-saving feature in Wi-Fi 6 that allows devices to schedule wake-up times to communicate with the router. This is particularly important for IoT ecosystems, where energy-efficient operation is crucial for battery-powered devices such as smart sensors and wearables. TWT extends battery life and reduces the need for frequent recharging, which is vital for Edge AI applications in remote or hard-to-maintain locations.
- **Enhanced Security:** With increasing concerns about data privacy, Wi-Fi 6 includes WPA3 encryption, offering stronger protection against unauthorized access. WPA3 enhances security by employing advanced cryptographic techniques to secure data transmission, which is especially important for Edge AI applications dealing with sensitive data, such as those in healthcare or finance.

Edge AI Suitability:

Wi-Fi 6 is ideally suited for Edge AI in scenarios involving heavy device communication that must be both effective and fast. In smart homes, Wi-Fi 6 can easily coordinate devices like voice assistants, security cameras, and smart thermostats for local processing and real-time responses. In industrial settings, Wi-Fi 6 enables real-time data transfer from sensors, robots, and AI-powered machines, facilitating quicker decision-making for automation, predictive maintenance, and optimization. Wi-Fi 6 also supports high device density in environments like offices and public spaces, where multiple devices can communicate at high speeds without network congestion. This is particularly beneficial for AI applications in workplace productivity tools, AI-based video conferencing, and collaborative work environments that require minimal latency in transferring data between devices and edge nodes.

In summary, Wi-Fi 6 offers high data throughput, efficient use of the radio spectrum, power savings for devices, and enhanced security—making it the ideal solution for Edge AI applications that demand reliable, low-latency connectivity in dense environments. As Edge AI continues to proliferate across industries, Wi-Fi 6 will play a crucial role in enabling efficient communication and data processing at the edge.

C. LPWAN: LOW-POWER, LONG-RANGE CONNECTIVITY

Low-power wide-area networks (LPWAN) refer to a class of wireless network technologies developed for applications requiring long-range communication with the lowest possible power consumption. LPWAN technologies, such as LoRaWAN, Sigfox, and NB-IoT, are perfect fits for Edge AI systems involving remotely dispersed devices in large areas, where the volume and frequency of data transmission needs are low [10].

LPWAN is increasingly used in fields such as agriculture, environmental monitoring, and asset tracking—areas where devices need to operate for extended periods without frequent recharging.

LPWAN Key Features for Edge AI:

- **Long Range:** LPWAN technologies are capable of providing coverage over several kilometers in rural settings and a couple of kilometers in urban environments. This extensive range makes LPWAN ideal for applications where edge devices are spread across large areas. For example, in agriculture, sensors deployed to monitor soil conditions, temperature, and crop health can transmit data over large distances, enabling continuous monitoring of expansive fields with minimal manual intervention.
- **Power Consumption:** One of the most notable features of LPWAN is its energy efficiency. These networks are designed to support battery-operated devices that can work for years without requiring frequent recharging or battery replacement. This feature is particularly valuable in remote areas where regular maintenance is challenging. Environmental sensors and tracking devices can operate autonomously for extended periods, reducing operational costs and ensuring the longevity of Edge AI applications in remote environments.
- **Scalability:** LPWAN can support large numbers of connected devices, making it an ideal solution for IoT deployments that require vast sensor networks. This scalability is essential for applications like environ-

mental sensing and smart agriculture, where hundreds or thousands of devices need to be deployed across large areas. LPWAN enables systems to grow efficiently without major infrastructure changes, accommodating the increasing data and connectivity demands of Edge AI.

Suitability for Edge AI:

LPWAN is particularly suitable for Edge AI systems that require extensive coverage and low power consumption. In agriculture, LPWAN can cover large fields with sensors that monitor soil moisture, weather conditions, and crop health. In environmental monitoring, devices that track air quality, temperature, and humidity can use LPWAN to transmit their data over vast distances. Asset tracking also benefits from LPWAN's ability to support low-power sensors over large distances, facilitating real-time inventory management, fleet tracking, and monitoring goods in transit. While LPWAN offers transfer rates lower than 5G or Wi-Fi 6, it is well-suited for transmitting small, periodic sensor data to edge devices or the cloud for processing.

VI. CHALLENGES IN WIRELESS FOR EDGE AI CONNECTIVITY

Wireless connectivity plays a crucial role in the performance of Edge AI systems, as continuous communication between distributed devices and edge servers is essential. However, several challenges hinder the efficiency, reliability, and scalability of wireless networks, potentially undermining the performance of various Edge AI applications. These challenges include network congestion, signal interference, coverage issues, latency, and security concerns.

A. NETWORK CONGESTION

One significant challenge is network congestion, particularly in environments with many devices connected to the same wireless network. As the number of connected devices increases, especially within IoT-heavy environments, the available bandwidth becomes stretched thin, leading to reduced data transmission. Network congestion results in increased latency and reduced throughput, which can be particularly problematic for real-time Edge AI applications.

For example, in smart cities or industrial IoT environments, congestion can delay the transmission of important data from numerous sensors, cameras, and other devices, which can have cascading effects on end-to-end decision-making processes.

B. INTERFERENCE IN SIGNALS

Signal interference is another critical issue in wireless communication for Edge AI. This interference can be caused by various factors, including physical obstacles, electromagnetic interference from other devices, and environmental elements like weather conditions. In dense urban environments, tall buildings can block signals, reducing connectivity quality. Performance can also be impacted by interference from other wireless signals, such as devices operating on the same or overlapping frequency bands (e.g., Wi-Fi or Bluetooth).

For Edge AI applications that require consistent and high-quality data transmission, such as autonomous vehicles or real-time video processing, any interference can lead to errors or delays in data transfer, compromising the accuracy and reliability of AI-driven decisions.

C. COVERAGE ISSUES

Coverage issues present another challenge, particularly in topographically remote or rural areas. Most radio technologies, such as Wi-Fi and LPWAN, inherently have limited range or poor coverage in such settings. For Edge AI systems deployed in agriculture, smart cities, or industrial facilities, ensuring network coverage across large or obstructed areas is crucial.

Coverage gaps can result in intermittent connectivity, which in turn affects continuous data intake and processing. In such cases, edge devices may need to rely on intermittent communications with the central network or cloud, reducing the efficiency of AI processing at the edge.

D. LATENCY AND RELIABILITY

Edge AI systems are designed to process data locally at the edge, with minimal latency, reducing reliance on cloud-based processing. However, wireless networks often face difficulties in delivering the low latency and high reliability required by real-time AI applications. For example, in autonomous vehicles, timely communication between sensors and processing units is critical; even slight delays or dropped packets can be unsafe. Similar quality-of-service (QoS) issues, such as high latency or poor network reliability, can affect applications like industrial automation or telemedicine, leading to costly mistakes.

E. SECURITY CONCERNS

Wireless networks are inherently more vulnerable to attacks and security breaches than wired networks. For Edge AI systems, where sensitive data may be transmitted over the air, it is essential to implement strong security protocols. Applications in fields like healthcare, finance, and smart cities are particularly susceptible to application-layer attacks, such as man-in-the-middle or denial-of-service attacks, which can compromise communication, data integrity, and personal information.

VII. WIRELESS NETWORK OPTIMIZATION TECHNIQUES FOR EDGE AI

Optimizing wireless networked systems is critical for the efficient deployment of Edge AI applications. These systems rely on real-time processing at the edge, making network efficiency essential to reduce latency while ensuring reliability and high throughput. Techniques such as adaptive beamforming, network slicing, interference mitigation, and load balancing are being explored to optimize wireless performance in meeting the high demands of Edge AI systems.

A. ADAPTIVE BEAMFORMING

Adaptive beamforming is a technology used in wireless communication to increase the sensitivity and quality of wireless signals. It achieves this by dynamically adjusting the phase and amplitude of signals from multiple antennas to focus the signal on specific receivers or groups of receivers. This method enhances the robustness of the signal, suppresses interference, and increases network efficiency, especially in environments with high device density or complex topography.

In the context of Edge AI, adaptive beamforming helps ensure that data transmission to and from edge devices, such as IoT sensors, cameras, and drones, remains steady and reliable. By focusing the wireless signal on the required area, communication quality is boosted, packet loss is reduced, and the accuracy of AI-driven decisions is greatly improved. For example, adaptive beamforming enhances the responsiveness of autonomous vehicles or robots, where constant data exchange with other entities is crucial for providing stronger and more reliable wireless connectivity.

B. NETWORK SLICING

Network slicing is a technique used in modern wireless networks, particularly in 5G, where multiple virtual, independent networks (slices) are created on top of a common physical network infrastructure. Each slice is tailored to meet the specific requirements of a given application, such as bandwidth, latency, or reliability, allowing for more efficient resource allocation. This ensures that time-sensitive applications in Edge AI are not interfered with by less critical traffic.

For instance, a network slice serving autonomous vehicles in smart cities would prioritize ultra-low latency and high reliability, while another slice serving environmental monitoring devices would focus on battery efficiency and range. Network slicing guarantees that different Edge AI applications perform optimally without compromising performance or quality of service (QoS).

Dynamic adjustments enable operators to realign network resources in real time to meet the evolving needs of Edge AI systems.

C. INTERFERENCE MITIGATION

Interference is a significant concern in wireless communication, particularly in high-density device environments. Sources of interference include other devices operating on the same or overlapping frequency bands, physical obstacles, and environmental factors such as weather conditions. Several techniques are available to reduce interference, including dynamic frequency selection, power control, and interference cancellation.

- **Dynamic Frequency Selection**: This technique allows systems to dynamically switch between frequency bands to avoid channels experiencing congestion or noise. By selecting less-interfered frequencies, Edge AI systems ensure reliable communication without performance degradation.
- **Power Control**: By adjusting transmission power, devices can reduce interference with other devices operating on the same frequency band. Edge AI systems can adjust power based on proximity and network conditions to maintain robust communication while minimizing interference.
- **Interference Cancellation**: Advanced algorithms are used to cancel unwanted signals and noise, particularly useful in environments with high interference, such as urban or industrial settings.

D. LOAD BALANCING AND TRAFFIC SHAPING

Load balancing and traffic shaping ensure that network traffic is evenly distributed over available channels, preventing bottlenecks and optimizing network performance. This approach is essential in Edge AI systems that generate large volumes of data, as it ensures minimal delay and maximizes system efficiency.

- **Load Balancing**: This technique distributes data across multiple communication paths, ensuring optimal resource utilization and avoiding congestion.
- **Traffic Shaping**: Traffic shaping identifies time-sensitive traffic, prioritizes it, and ensures that applications requiring low latency—such as autonomous vehicles or remote surgery—are not impacted by less urgent data transmission.

VIII. SCALABILITY IN WIRELESS NETWORKS FOR APPLICATIONS RELATED TO EDGE AI

As the number of connected edge devices and AI-powered applications increases, scalability becomes a critical factor in the design of wireless networks. When demand for real-time data processing and communication rises, scalability is essential to ensure that wireless networks can support the increased load.

Edge AI, which involves handling various devices and sensors at the edge of the network for data collection and processing, is a key enabler for scalable wireless networks. These networks support large-scale AI applications by enabling efficient operation with ease. Scalability in wireless networks for Edge AI is achieved through various factors and technologies, including network architecture, resource management, and advanced technologies such as 5G, Wi-Fi 6, and LPWAN.

A. NETWORK ARCHITECTURE

The deployment of large-scale Edge AI applications requires scalability in the underlying wireless network architecture. Traditional wireless network models centralize processing in a cloud or data center, which limits scalability and often introduces latency due to bottlenecks when managing a large number of devices.

In contrast, Edge AI systems leverage decentralized architectures, where processing is distributed closer to the devices, reducing the load on central systems and improving scalability. Edge AI deployments benefit from hierarchical

or multi-tiered network architectures, where data processing occurs at the edge of the network, near the source of the data.

This distribution of computational tasks across edge devices, gateways, and local servers allows for scalability as the number of connected devices grows, without overloading central servers. As a result, this approach reduces network congestion, latency, and facilitates real-time processing for AI data.

B. RESOURCE MANAGEMENT AND DYNAMIC ALLOCATION

Efficient resource management is crucial for scalability in wireless networks supporting Edge AI. As more devices connect to the network, dynamic adjustments in bandwidth, processing power, and other resources are necessary to maintain seamless communication and prevent interference or delays.

Network slicing is one effective technique used to manage resources more efficiently by dividing the network into virtual slices tailored to specific application needs. Each slice can provide the required bandwidth and latency for different Edge AI use cases, such as autonomous vehicles or smart cities. Dynamic and flexible resource allocation is vital to ensure high performance and scalability in Edge AI applications.

C. ADVANCED WIRELESS TECHNOLOGIES

Advanced wireless technologies, such as 5G, Wi-Fi 6, and LPWAN, are key enablers of scalability in wireless networks. These technologies are designed to support a large number of connected devices while maintaining stable and efficient communication.

- **5G**: 5G networks provide ultra-low latency and massive device connectivity, which is essential for scalable Edge AI systems. With a capacity of up to 1 million devices per square kilometer, 5G supports large-scale IoT networks, such as smart cities or industrial automation systems, without performance degradation.
- **Wi-Fi 6**: Wi-Fi 6 improves scalability in high-density environments by enhancing throughput and reducing congestion. It employs techniques like OFDMA and TWT to optimize traffic management and efficiency in high-density application environments.
- **LPWAN**: Low Power Wide Area Networks (LPWAN), such as LoRaWAN, are designed to support thousands of devices over long distances with minimal power consumption. While LPWAN offers lower data transfer rates than other wireless technologies, it is ideal for Edge AI use cases that require small data volumes over wide areas, such as environmental monitoring or smart agriculture.

D. EDGE COMPUTING AND DATA LOCALIZATION

Edge computing enables the scalability of wireless networks for Edge AI applications by moving data processing closer to the source of data generation. By relocating computational tasks from the cloud to edge devices or local data centers, edge computing reduces latency and bandwidth usage at the edge.

This approach supports scalable Edge AI applications by enabling them to operate independently of the cloud. Distributed computing and storage at the edge further reduce backhaul communications, which can become a bottleneck as the number of devices increases. Only the data of interest is sent to the cloud or central systems, minimizing the load on the network and ensuring efficient operation.

IX. LATENCY AND QOS IN WIRELESS NETWORK FOR EDGE AI

Latency reduction and Quality of Service (QoS) are critical factors in ensuring real-time performance for AI models running at the edge. Edge AI refers to a topology where AI models are deployed on devices or edge nodes, such as IoT devices, local servers, and gateways, that process data close to the source rather than relying on a centralized cloud server.

These AI models at the edge must meet real-time decision-making demands with low-latency connectivity and a high standard of QoS. Addressing these factors ensures timely responses, which is essential for applications such as autonomous vehicles, industrial automation, and smart healthcare [9].

A. ROLE OF LATENCY IN EDGE AI

Latency refers to the delay between data generation by an edge device and the corresponding action or response. In Edge AI applications, minimizing latency is crucial to enable near real-time processing and decision-making. For instance, AI models operating on sensor data like LIDAR and cameras in autonomous vehicles must make split-second decisions regarding navigation and collision avoidance. Similarly, in industrial automation, latency in AI-driven actions can result in errors, safety hazards, or inefficiencies.

Wireless networks designed to support Edge AI must provide ultra-low latency to meet the requirements of time-critical applications. Technologies like 5G and Wi-Fi 6 are engineered to reduce latency, with 5G theoretically achieving latencies as low as 1 ms, making it an ideal choice for real-time AI applications. Furthermore, edge computing reduces latency by processing data closer to the source, eliminating the need to transport data to the cloud for processing, thereby avoiding significant delays.

B. QUALITY OF SERVICE (QOS) IMPORTANCE

QoS refers to the network's ability to guarantee a specific level of performance, especially under heavy traffic or complex application scenarios. In Edge AI applications, consistent and predictable network performance is vital to ensure that AI models can function effectively.

QoS management involves prioritizing specific types of traffic, minimizing packet loss, and ensuring reliable data transmission. For example, AI-based video surveillance or

real-time video processing requires high-bandwidth and low-latency connectivity to transmit clear image and video feeds continuously. Without QoS, network congestion or packet loss could degrade AI performance, resulting in incorrect predictions or system failures.

Technologies like network slicing and traffic shaping play a critical role in ensuring QoS in wireless networks. Network slicing enables the creation of multiple virtual networks on top of a single physical network, each tailored to specific application requirements with guaranteed bandwidth, latency, and reliability. For example, Edge AI applications in autonomous vehicles may have dedicated slices with stringent latency and bandwidth constraints, while less time-critical IoT applications may use less demanding slices. Traffic shaping helps manage data flow by prioritizing critical AI data, ensuring uninterrupted communication.

C. LATENCY AND QOS IN 5G, WI-FI 6, AND LPWAN

5G technology, characterized by ultra-low latency, supports high-priority traffic management, making it well-suited for Edge AI systems operating in dynamic environments. Wi-Fi 6 introduces features like Orthogonal Frequency-Division Multiple Access (OFDMA) and Target Wake Time (TWT), which optimize bandwidth allocation and reduce congestion, improving QoS in high-density networks.

Low-Power Wide-Area Networks (LPWAN), such as LoRaWAN, cater to low-bandwidth and long-range scenarios where latency is less critical. While LPWAN technologies may not meet the real-time performance requirements of some Edge AI applications, they are ideal for use cases like environmental monitoring and smart agriculture that involve broad coverage and minimal data transfer [4].

D. CONCLUSION

In Edge AI applications, minimizing latency and ensuring high QoS are paramount for real-time decision-making and reliable system performance. Advanced wireless technologies such as 5G and Wi-Fi 6, along with techniques like network slicing and traffic shaping, enable wireless networks to meet the demanding requirements of Edge AI systems. These technologies collectively support the scalability and efficiency needed for next-generation AI-driven applications.

X. FUTURE TRENDS IN WIRELESS NETWORKS FOR EDGE AI

As Edge AI continues to evolve, advancements in wireless networking are keeping pace with increasing demands for speed, reliability, and scalability. Innovations such as 6G, satellite internet, and AI-driven network management are poised to revolutionize wireless connectivity, enabling Edge AI applications to achieve unprecedented efficiency and intelligence. These emerging trends aim to address current bottlenecks and expand the potential of edge-based intelligence across diverse industries.

FIGURE 1. System Architecture [9]

A. 6G: THE NEXT FRONTIER OF WIRELESS CONNECTIVITY

Building on the capabilities of 5G, 6G is expected to introduce groundbreaking advancements in wireless networking, including ultra-high speeds, near-zero latency, and significantly enhanced spectrum efficiency. Anticipated to be deployed by the 2030s, 6G networks aim to achieve data rates of up to 1 Tbps and latencies as low as 100 microseconds.

Key features of 6G that will benefit Edge AI include:

- **Integrated Sensing and Communication:** Combining data transmission with environmental sensing will enable context-aware AI for advanced robotics and precision healthcare applications.
- **Terahertz Frequencies:** Leveraging higher frequency bands will allow for extremely high data throughput, supporting applications like real-time digital twins and holographic communications.
- **AI-Native Architecture:** AI will be integrated into every layer of 6G networks, facilitating optimal resource allocation, traffic management, and energy efficiency.

These advancements will empower Edge AI systems to handle increasingly complex workloads in real-time, enabling futuristic applications such as fully autonomous vehicles and immersive augmented and virtual reality experiences.

B. SATELLITE INTERNET FOR GLOBAL COVERAGE

Satellite internet, driven by initiatives like Starlink and OneWeb, holds immense promise for Edge AI, particularly in rural or underserved regions. By extending connectivity to areas beyond the reach of traditional networks, satellite internet enables Edge AI applications in diverse fields such as:

- **Agriculture:** Facilitating remote crop and livestock monitoring.
- **Disaster Management:** Enabling real-time data analysis to enhance emergency response.
- **Environmental Monitoring:** Supporting Edge AI-powered sensors for climate studies.

The global coverage offered by satellite internet eliminates geographical constraints, fostering inclusivity and innovation in remote areas, and enabling the widespread deployment of Edge AI systems.

C. AI-DRIVEN NETWORK MANAGEMENT

AI is increasingly pivotal in managing the complexities of wireless networks. AI-driven network management leverages machine learning models for dynamic monitoring, analysis, and optimization of network performance. For Edge AI, this translates into:

- **Adaptive Resource Allocation:** Dynamically optimizing bandwidth and latency for critical applications.
- **Proactive Maintenance:** Predicting and addressing network failures before they occur.
- **Traffic Optimization:** Reducing congestion and interference to ensure smooth data flow.

AI-enhanced networks will enable self-healing and self-optimization capabilities, ensuring consistent performance under the highly variable conditions of different Edge AI applications.

XI. APPLICATIONS OF EDGE AI

Edge AI has been widely adopted across industries for real-time decision-making, low latency, and localized processing. This section explores key use cases and the impactful ways AI enhances edge computing across various domains.

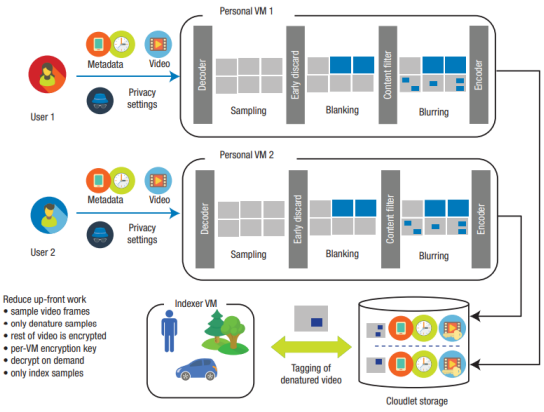

FIGURE 2. GigaSight framework. A cloudlet performs computer vision analytics on video from mobile devices in near real time and only sends the results along with metadata to the cloud, sharply reducing ingress bandwidth into the cloud. VM: Virtual machine [2].

A. HEALTHCARE

Edge AI plays a critical role in the healthcare sector by enabling the rapid processing of massive amounts of data generated by medical devices, wearables, and IoT sensors. This facilitates real-time insights while maintaining data privacy.

Key Applications:

- **Remote Patient Monitoring:** Wearable devices and sensors capture vital signs like heart rate, oxygen levels, and glucose levels. Edge AI processes this data locally, enabling real-time monitoring and analysis for patient care.
- **Medical Imaging:** AI models deployed on edge devices analyze medical images (e.g., X-rays, MRIs) quickly, reducing latency and improving diagnostic efficiency in critical scenarios.

B. AUTONOMOUS VEHICLES

Autonomous vehicles rely heavily on edge AI for real-time decision-making by processing sensor, camera, and radar data locally. Cloud-based processing is impractical due to latency constraints.

Key Applications:

- **Real-time Decision Making:** Edge AI processes sensor data to make split-second decisions, such as detecting obstacles or navigating through dynamic environments.

C. SMART CITIES

Edge AI is integral to smart cities, improving urban infrastructure and enabling smarter, more efficient operations in traffic management, public safety, and energy distribution.

Key Applications:

- **Traffic Management:** Edge AI monitors traffic patterns, adjusts signals dynamically, and detects congestion or accidents in real-time.
- **Public Safety:** Surveillance cameras with edge AI detect suspicious activities and provide real-time responses, ensuring security without transmitting large video data to central servers.

D. INDUSTRIAL INTERNET OF THINGS (IIOT)

Edge AI transforms industrial operations by enabling real-time analytics for predictive maintenance and operational optimization.

Key Applications:

- **Predictive Maintenance:** Edge AI monitors machinery performance to detect anomalies early, minimizing downtime and extending equipment life.
- **Operational Optimization:** Factories use edge AI to monitor production, correct inefficiencies, and make real-time adjustments, enhancing productivity and reducing waste.

E. RETAIL AND E-COMMERCE

Retail and e-commerce industries leverage edge AI to optimize customer experiences and streamline operations.

Key Applications:

- **Personalized Shopping Experiences:** Edge AI analyzes customer behavior to provide real-time recommendations without relying on cloud processing.
- **Smart Inventory Management:** Retailers use edge AI to track inventory levels, predict demand patterns, and optimize stock management [3].

XII. CONCLUSION

The integration of artificial intelligence with edge computing offers unparalleled advantages in processing and utilizing data across industries. Edge AI addresses traditional cloud computing limitations, such as latency, bandwidth

constraints, and data privacy concerns, by bringing computational power closer to the source of data generation.

In this paper, we explored the challenges of deploying AI in edge environments, including resource constraints, energy efficiency, scalability, and security. Solutions like model compression, federated learning, and specialized hardware development have been proposed to overcome these limitations.

Real-world applications demonstrate the transformative potential of edge AI, from real-time patient monitoring in healthcare to split-second decision-making in autonomous vehicles. Edge AI is revolutionizing urban infrastructure through smart city initiatives, improving industrial efficiency via IIoT, and enhancing retail experiences.

With the expansion of 5G networks and the proliferation of IoT devices, the possibilities of edge AI continue to grow. Despite ongoing challenges, such as ensuring data privacy and improving energy efficiency, the future of edge AI is promising, driving intelligent, adaptive, and secure systems for real-time applications.

XIII. CONTRIBUTIONS

The research investigates the integration of edge computing and AI, showing the capability of this combination to overcome some of the traditional challenges with cloud computing, including high latency and bandwidth limits.

Sai Varshith Gandu provided a comprehensive overview of the role of wireless networks in enabling edge AI and exploring the architectures for supporting edge AI. Further, his contribution covered features and suitability of technologies such as 5G, Wi-Fi 6, and LPWAN about Edge AI systems.

Tarun Talasila focused on the challenges of deploying AI at the edge, with a particular emphasis on wireless connectivity. His view ranged from network congestion and signal interference down to coverage issues, which are very important in the case of wireless networks for edge AI. He introduced some optimization methods, including those using adaptive beamforming and interference mitigation techniques.

Balaji Chappidi spoke about the scalability of wireless networks concerning Edge AI applications. He further spoke about how wireless networks need to cater to a large number of connected devices. Latency reduction and maintenance of QoS within wireless networks were also reviewed as very important parameters that ensure real-time performance of edge AI models.

Sowgoto Raha Sunny presented future trends in wireless networking for Edge AI, including emerging technologies on 6G, satellite Internet, and AI network management. She showed that those aspects extend to the next generation of wireless, presenting sophisticated and reliable edge AI applications enabled.

REFERENCES

[1] H. Li, Y. Dai, and C. Yi, "Private 5G Networks: Architecture, Technologies, and Security Aspects," in IEEE Explore, 2020.

[2] M. Satyanarayanan, "The Emergence of Edge Computing," Computer, vol. 50, no. 1, pp. 30-39, Jan. 2017.

[3] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., Arcas, B. A. y. (2017). "Communication-efficient learning of deep networks from decentralized data." Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS).

[4] F. Wang, M. Zhang, X. Wang, X. Ma, and J. Liu, "Deep Learning for Edge Computing Applications: A State-of-the-Art Survey," in IEEE Access, vol. 8, pp. 58322-58336, 2020. doi: 10.1109/ACCESS.2020.2982411.

[5] Y. Abbas, Y. Zhang, A. Taherkordi, and T. Skeie, "Mobile Edge Computing: A Survey," IEEE Internet of Things Journal, vol. 5, no. 1, pp. 450-465, Feb. 2018.

[6] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge Computing: Vision and Challenges," in IEEE Internet of Things Journal, vol. 3, no. 5, pp. 637-646, Oct. 2016. doi: 10.1109/JIOT.2016.2579198.

[7] "Table of Contents," in IEEE Industry Applications Magazine, vol. 24, no. 5, pp. 1-2, Sept.-Oct. 2018. doi: 10.1109/MIAS.2017.2781027.

[8] O. Y. Fai, Y. V. Voon, and H. Nisar, "Image classification using wavelet coefficients for video surveillance," in IEEE Conference on Systems, Process and Control, Melaka, Malaysia, 2016, pp. 107-112. doi: 10.1109/SPC.2016.7920713.

[9] S. Wang, et al., "Adaptive Federated Learning in Resource Constrained Edge Computing Systems," in IEEE Journal on Selected Areas in Communications, vol. 37, no. 6, pp. 1205-1221, June 2019. doi: 10.1109/JSAC.2019.2904348.

[10] Y. Liu, H. Xu, and W. C. Lau, "Online Job Scheduling with Resource Packing on a Cluster of Heterogeneous Servers," in IEEE INFOCOM 2019, Paris, France, 2019, pp. 1441-1449. doi: 10.1109/INFOCOM.2019.8737465.