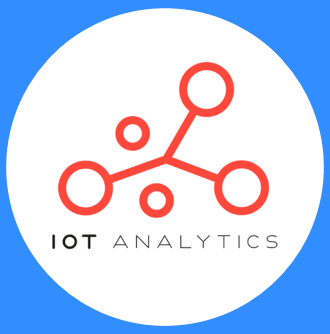
ANALYST REPORT

The evolution of AI-driven IoT: Trends, innovations and market growth in cellular connectivity



SPONSORED BY

Fibocom

AI-driven IoT and cellular connectivity

Cellular IoT connections are to outpace overall IoT connections through to 2030. By the end of 2024, IoT Analytics estimates the number of connected IoT devices reached 18.8 billion, approximately a 13% increase YoY. By the end of 2030, these connected IoT devices are projected to grow at 14% CAGR, surpassing 40 billion, writes Satyajit Sinha, a principal analyst at IoT Analytics.

Of these connected IoT devices, approximately 22% were cellular IoT connections, reaching 4.2 billion by the end of 2024. Cellular IoT connections are expected to grow at 15% CAGR, reaching 9.7 billion by the end of 2030.

Satyajit Sinha
Principal Analyst

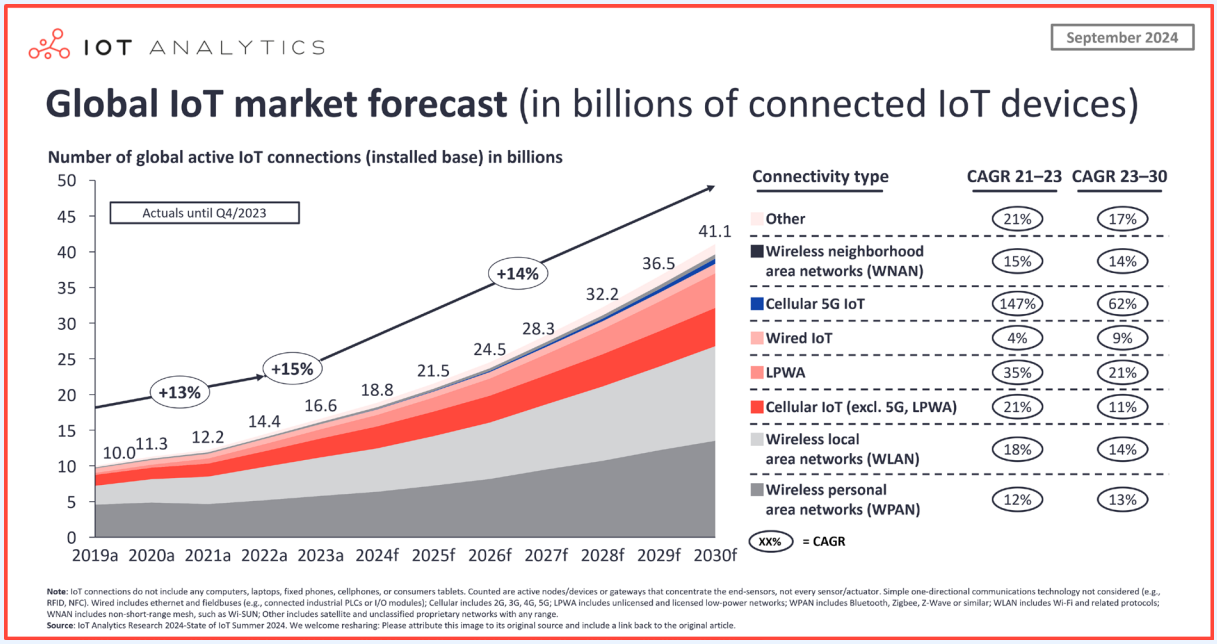


Figure 1: Forecast for connected IoT devices

Source: IoT Analytics

Key IoT connectivity trends

Four key trends are driving this growth in cellular IoT connections.

Trend 1: Fixed wireless access and the automotive sector are driving 5G IoT growth

Fixed wireless access (FWA) dominated 5G IoT connections in 2024. FWA contributed 46% of global 5G IoT connections in 2024 by utilising public 5G networks to deliver broadband internet services delivered wirelessly to a fixed location like homes and businesses, particularly in areas lacking fibre infrastructure. This approach offers a cost-effective alternative to traditional broadband services and expands connectivity in underserved regions.

The automotive sector accounts for a quarter of the global 5G IoT connections in 2024. The automotive sector, including transportation, supply chain and logistics, accounted for 26% of global 5G IoT connections in 2024, propelled by connected vehicles. The industry is integrating 5G IoT to enhance real-time navigation, telematics and infotainment. Additionally, 5G supports future technologies like cellular vehicle-to-everything (C-V2X), enabling communication between vehicles, infrastructure and other road users.

5G technologies like URLLC, mmWave and MIMO ensure long-term compatibility for C-V2X communication. 5G's ultra-reliable low-latency communications (URLLC) capabilities enable sub-millisecond latency, essential for real-time C-V2X interactions like collision avoidance and autonomous driving. While mmWave offers high data rates, sub-6 GHz bands allow for reliable communication in safety-critical scenarios. Massive multiple-input and multiple-output (MIMO) technology improves network efficiency, complemented by direct sidelink communication. Meanwhile, deploying edge servers reduces delays by bringing data processing closer to the vehicles, and 5G SA network slicing enables dedicated, optimised network partitions for automotive applications. With vehicles remaining in use for over a decade, integrating 5G ensures long-term compatibility with emerging technologies.

"[FWA] is a technology that provides broadband internet access to homes and businesses using 4G and 5G cellular networks instead of traditional wired connections like fibre or cable. The key advantages for telecoms operators are threefold: rapid deployment, cost-effectiveness and scalability. This means that FWA can be quickly rolled out in regions lacking wired infrastructure, it's more affordable — especially in rural and underserved areas — and it can be easily expanded as demand grows. For users, FWA delivers high-speed connectivity comparable to wired networks, making it ideal for activities like streaming, gaming, and remote work."
Paul Anand, sales director for EMEA at Fibocom ▶

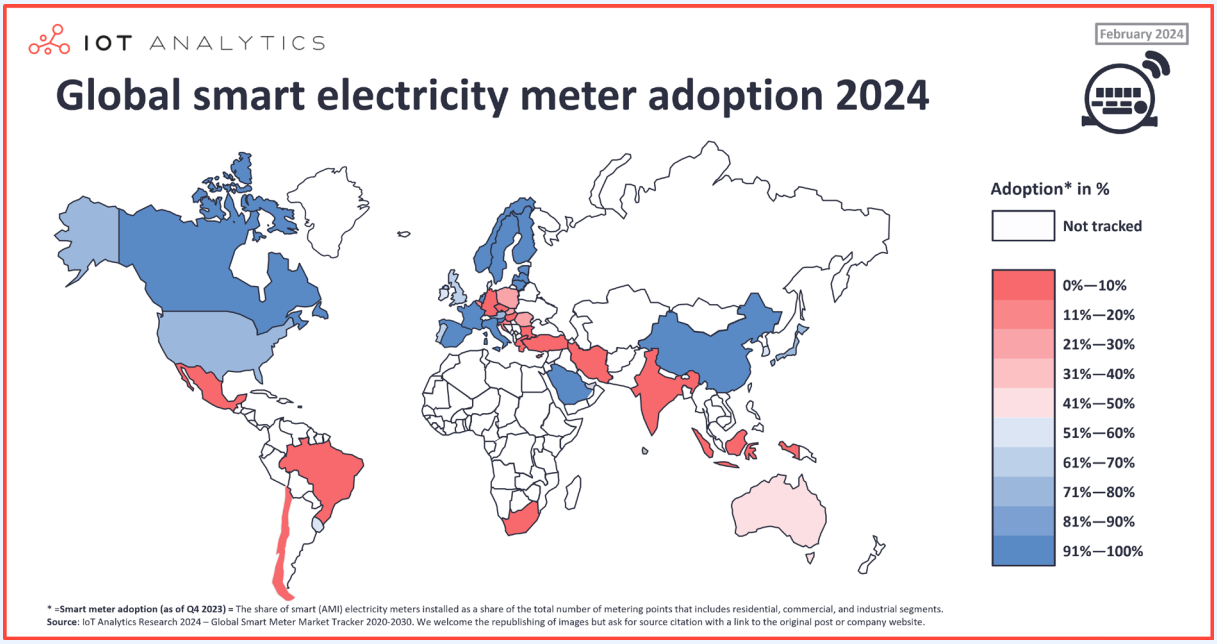


Figure 2: Smart electricity meter adoption by country

Source: IoT Analytics

Trend 2: The evolution of streamlined IoT connectivity with 5G RedCap

5G reduced capability (RedCap) to drive IoT connectivity growth by 2030. RedCap is expected to grow at 18% CAGR from 2025 to 2030 and account for 6% of mobile operator IoT revenue by the end of this period despite contributing to only 2% of the total global cellular IoT connections. This growth highlights RedCap's role in supporting cost-efficient IoT applications that do not require high-performance parameters similar to mobile broadband.

Optimised performance requirements to support diverse IoT applications. Bridging the gap between massive IoT and high-performance 5G, RedCap was designed to support use cases whose requirements fall between the more extreme requirements defined for massive machine type communications (mMTC), enhanced mobile broadband (eMBB) and URLLC. It supports data rates below 150 Mbps and latency under 100 milliseconds, reducing device complexity from a design perspective and associated costs. Further, its optimised design enables broad adoption across consumer, enterprise and industrial sectors — from wearables to industrial sensors and smart grids — with video surveillance emerging as a notable growth area since RedCap provides sufficient uplink capacity to transmit high-quality video streams without the higher costs associated with standard 5G.

RedCap examples

- At MWC 2024, **Telit Cinterion** showcased a video surveillance application utilising their FN920C04 5G RedCap module within a **Digi International** router, which **Nokia**'s RAN and core network supported. 5G RedCap is ideal for mid-speed IoT applications like video surveillance and provides an improved uplink data rate compared to 4G with reduced device complexity.

- **Ericsson** also demonstrated a similar use case featuring 5G AIoT cameras from **Four-Faith**, integrated with 5G RedCap modules. This setup delivered a stable uplink capacity of 8 Mbps over their equipment, which surpasses the performance of LTE Cat-4.

"RedCap is a technology that bridges the gap between low power area protocols like CAT-M and narrowband IoT on one side, and high-performance 5G on the other. It was created to offer a balance between low latency, high bandwidth capabilities, and the ability to connect a large number of devices. This makes it perfect for applications in smart metering, industrial automation, wearables and even low-cost 5G fixed wireless access in emerging markets."
David Palmer, technical director at Fibocom ▶

Trend 3: Smart meters enabling smarter, greener utility solutions

Smart meter deployments are expected to surpass 1.75 million by 2030. By the end of 2024, the number of electricity, gas, and water smart meters deployed worldwide is estimated to have exceeded 1.1 billion, according to IoT Analytics' Global Smart Meter Market Tracker 2020–2030. This estimate excludes automated meter reading (or AMR) systems, as they lack two-way communication that would enable real-time data collection to help utilities optimise resource delivery (a key feature of smart meters). Driven by the expanding deployment of advanced metering infrastructure (AMI), sustainability and digital transformation, the number of smart meter deployments is forecasted to grow at 6% CAGR between 2024 and 2030, surpassing 1.75 million by the end of the decade.

India represents strong market opportunities. A key growth market to watch is India, where the government is undergoing a project to deploy 250 million smart meters by 2027. This initiative represents a US$20 billion opportunity in energy management to enhance operational efficiency, enable advanced data analytics, and optimise resource utilisation in the world's most populated country. Such large-scale projects that adopt cellular IoT technologies will play a crucial role in modernising energy.

"The smart meters bring a lot of benefits to the table for consumers. They provide real-time data on energy usage — eliminating the need for manual meter reading and reducing errors and disputes. This transparency not only builds trust with customers but also enables utilities to pinpoint peak usage times, optimise energy distribution, detect power theft and reduce technical losses. In essence, smart meters are a win-win, enhancing both grid efficiency and the overall reliability of energy delivery."
Ragin Kallanmar Thodikai, country sales manager for India at Fibocom Wireless

Trend 4: Cellular IoT technologies are evolving with edge-AI adoption

AI adoption in cellular IoT chipsets is accelerating. IoT connectivity chipset and module manufacturers are increasingly integrating AI accelerators like GPUs and NPUs into their products, especially in cellular IoT chipsets and modules. In 2023, AI-enabled cellular IoT modules comprised only 2% of total cellular IoT module shipments globally; however, by 2027, this share is expected to grow to 9%, according to IoT Analytics' Cellular IoT Module and Chipset Market Tracker.

AI accelerators enhance IoT chip efficiency and responsiveness. Integrating AI accelerators into cellular IoT chipsets offers benefits like real-time data processing with reduced latency, enabling on-device analysis and decision-making. Additionally, this integration provides greater efficiency and reduced device size, as connectivity modules are embedded, streamlining form factors. This shift is driving smarter, more autonomous and more responsive IoT solutions in areas like industrial IoT (IIoT) and enabling edge AI.

Deep-dive: Cellular edge AI

IoT is shifting from connectivity to real-time intelligence. While IoT connections are growing rapidly, IoT is evolving beyond connectivity to include data processing and real-time decision-making. Edge computing and AI accelerators are key to this evolution, bringing AI inference closer to data sources. By embedding these accelerators in edge devices — whether on a PCB or within a processor — latency and bandwidth bottlenecks are minimised, enabling instant decisions for critical applications like autonomous systems, industrial robotics and anomaly detection.

AI accelerators at different levels of edge

AI accelerators are being integrated across different layers of the edge computing hierarchy, particularly thick edge, thin edge and micro-edge, with each layer serving specific applications based on computational needs and proximity to data sources.

Edge computing brings intelligence closer to data sources. To understand how AI is being integrated at the edge, it is helpful to understand edge computing and its hierarchy. IoT Analytics defines edge computing as intelligent computational resources located close to the source of data consumption or generation. The edge includes all computational resources at or below the cell tower data centre and/or on-premises data centre, and there are three types of edges — thick, thin and micro — as shown in **Figure 3**. ▶

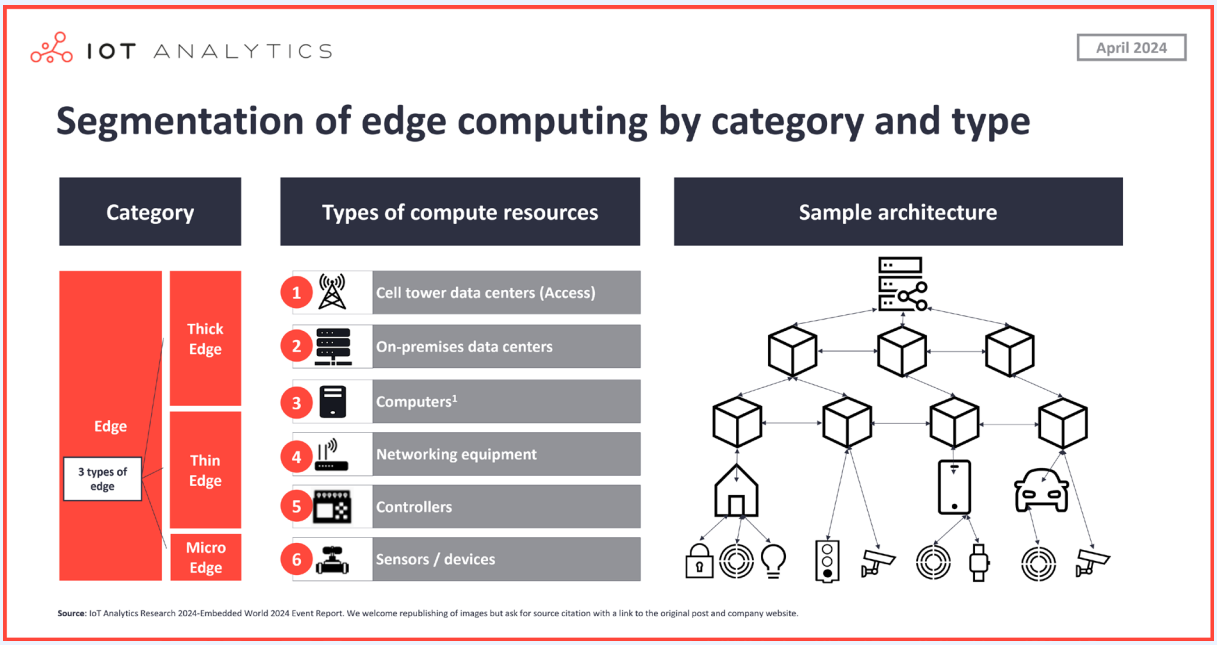


Figure 3: Layout of edge computing hierarchy

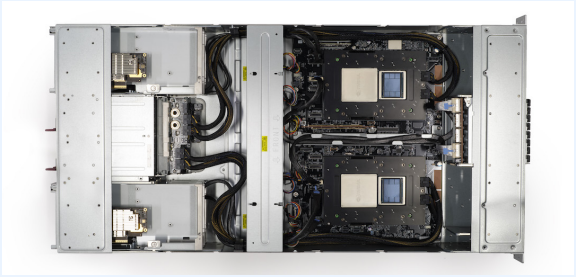
Source: IoT Analytics

Thick edge AI integration: Powering AI training and inference at the network edge

Thick edge computing reduces cloud reliance and enhances privacy. Thick edge computing uses high-performance computing resources like micro data centres or edge servers with powerful chipsets (CPUs, GPUs and NPUs) to handle compute-intensive tasks such as AI inference and localised model training. This shift from centralised cloud environments to thick-edge environments reduces cloud dependency, lowers operational costs and improves data privacy. By processing AI models on-premises or within vendor-managed edge infrastructure, businesses can optimiseresources while enabling fast, localised decision-making.

Thick edge AI example

HPE, a US-based information and server technology provider, offers its ProLiant Compute DL384 AI servers equipped with **NVIDIA**'s GH200 NVL2 GPUs, which support large language models and generative AI applications, including AI inference and fine-tuning for models like retrieval augmented generation.



HPE's Proliant Compute servers with NVIDIA GH200 NVL2 GPUs
(Source: NVIDIA)

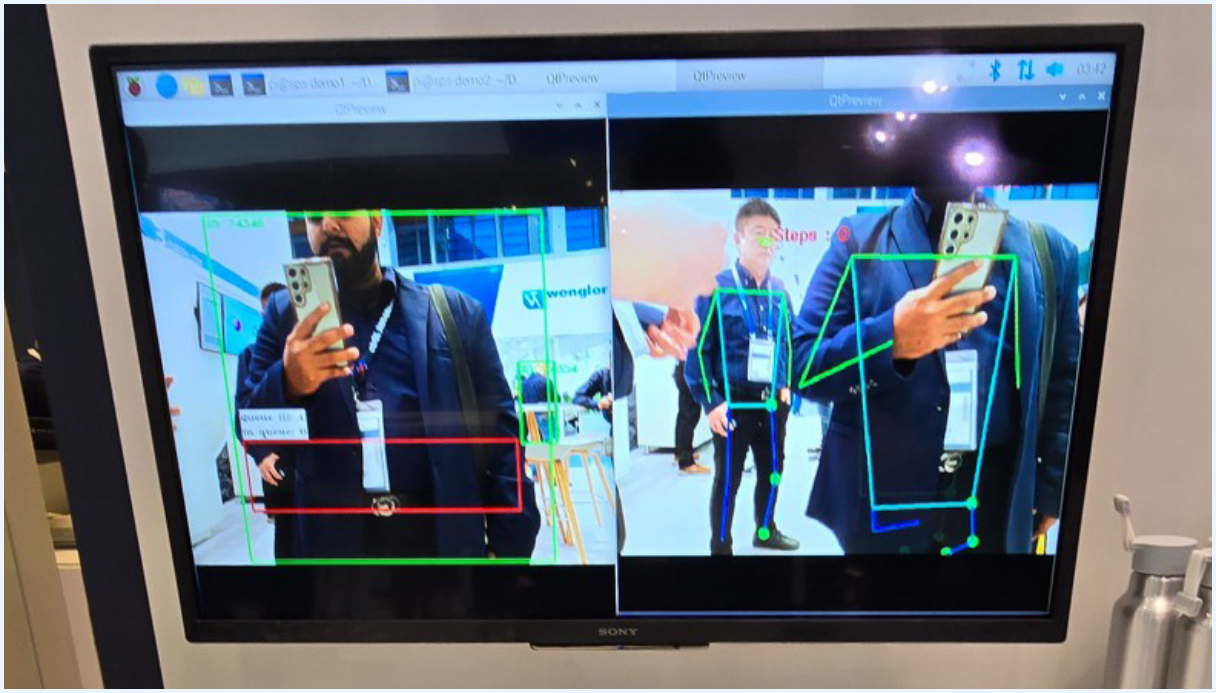
Thin edge AI integration: Enabling real-time intelligence at the IPC and IoT gateway level

Thin edge computing optimises data processing and network efficiency. Thin edge computing operates at the industrial PC (IPC) and IoT gateway level, where intelligent controllers and edge devices preprocess sensor data and aid decision-making. Thin edge devices connect to end nodes using wired, for example, Ethernet or wireless, for example, Wi-Fi or short-range protocol, standards and enable efficient data transmission to edge servers or cloud data centres, optimising real-time analytics while reducing cloud dependency. This architecture enhances network efficiency, automation and scalability, making it essential for industrial and enterprise environments.

The demand for real-time data processing and autonomous decision-making at the edge is growing, and hardware manufacturers are integrating advanced AI accelerators into IPCs and IoT gateways to meet this need. Integrating localised processing into these thin-edge devices reduces reliance on cloud or central servers, minimising latency and optimising operational efficiency.

Thin edge AI example

Eurotech, an Italy-based edge computing and industrial IoT solutions provider, offers ReliaCOR 33-11, a fanless IPC/IoT gateway and embedded edge AI system powered by the NVIDIA Jetson AGX Orin that supports AI applications in robotics, industrial automation and smart cities. It delivers up to 275 TOPS of AI performance due to its Ampere GPU with ▶



IMX500 embedded with Raspberry PI implementing AI-based machine vision
(Source: IoT Analytics)

2048 CUDA cores and 64 Tensor cores. The system includes connectivity options like LTE-Cat 4, Wi-Fi and Bluetooth.

Eurotech's ReliaCOR 33-11
(Source: Eurotech)

Micro-edge AI: Enabling intelligence at the sensor level

Micro-edge AI enables autonomous, intelligent device decision-making. Micro-edge AI computing integrates AI capabilities directly at the data-generation level, enabling intelligent sensors and devices to make autonomous decisions and improve scalability. As with the other levels, these devices are embedded with AI accelerators. Functioning as end nodes, micro-edge devices connect to host or master systems using wired and wireless communication standards.

Micro-edge AI example
Sony, a Japan-based electronics and information technology provider, integrated its IMX500 sensor with Raspberry Pi. The IMX500 is a CMOS sensor with built-in image processing and AI capabilities, allowing edge AI inference in a compact setup.

By integrating it with Raspberry Pi, Sony reduces infrastructure requirements and enables faster deployment of edge AI solutions, making the system more efficient for industrial applications.

Different AI-enabled chipsets categorised by computing power

AI-enabled IoT chipsets enhance connectivity and autonomy. As noted, IoT connectivity chipset and module manufacturers are increasingly embedding AI capabilities into their chipsets and modules — across both Wi-Fi and cellular IoT chipsets. Edge devices equipped with these benefit not only from more reliable connectivity and device autonomy but also additional embedded, scalable computing power, supporting diverse applications in industrial settings, such as in robotics for tasks like 3D mapping and autonomous mobile robot navigation (more on this later).

This power ranges from 0.2 TOPS to 48 TOPS, and IoT Analytics classifies this range into three categories: low capability, medium capability, and high capability:

1. **Low AI capability (less than 5 TOPS):** These modules are used for basic AI tasks like acoustic event detection, gesture recognition and voice ID. They accounted for 45% of global AI-enabled cellular IoT module shipments in 2024. Shipments of these modules are projected to grow at a CAGR of 41% until 2027, but modules with medium and high AI capabilities are expected to grow faster. ▶

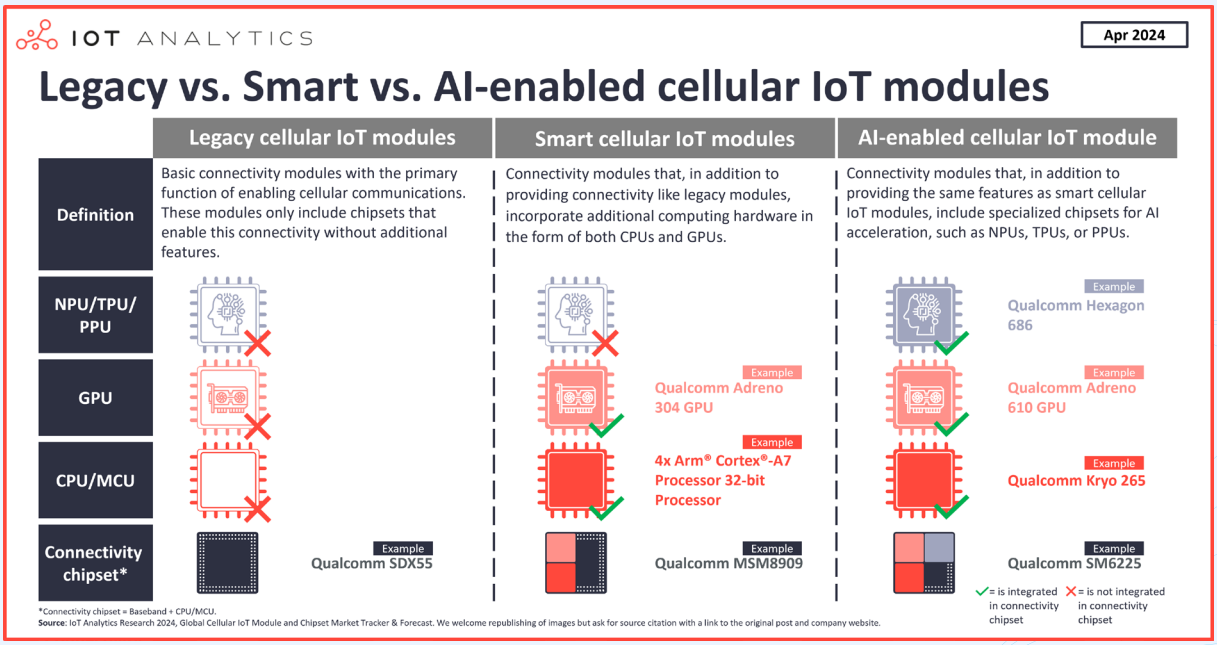


Figure 4: Legacy vs. smart vs. AI-enabled cellular IoT models and their integrated chipsets

Source: IoT Analytics

Example: *China-based wireless communication modules provider* ***Fibocom****'s SC138-EAU module with a* ***Qualcomm*** *QCM6125 SoC, offering 1 TOPS of AI performance*

2. **Medium AI capability (5–10 TOPS):** These modules handle more complex tasks such as human and vehicle detection, face recognition and people counting. They represented 43% of global AI-enabled cellular IoT module shipments in 2024, with a projected CAGR of 117% until 2027.

Example: *China-based IoT module provider* ***Quectel****'s SG-530C-CN module, featuring a UNISOC P778 SoC with 8 TOPS*

3. **High AI capability (Over 10 TOPS):** These modules support advanced applications like predictive maintenance, machine vision, driver safety solutions and intelligent voice assistance. They comprised 12% of global AI-enabled IoT module shipments in 2024, with shipments expected to grow at a CAGR of 97% until 2027.

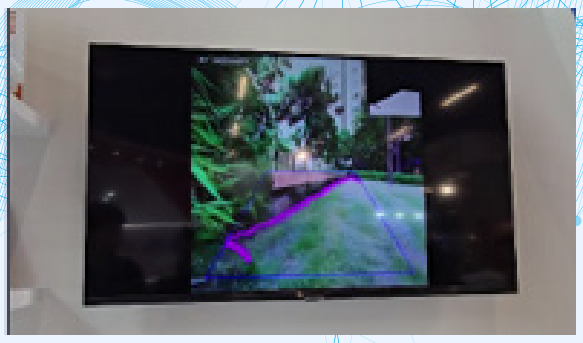
Example: *Fibocom's SCA825-W module with Qualcomm QCM825 and a 15 TOPS AI engine*

Chipset providers integrating AI into their offerings. Leading chipset providers like US-based Qualcomm have embedded AI acceleration in their connectivity solutions, offering pre-optimised hardware and an AI software ecosystem, such as the Qualcomm AI Hub, to streamline development. Other chipmakers like China-based **UNISOC** and Israel-based **Sony Semiconductor** are also integrating AI into their connectivity modules, driving a wider industry trend towards AI-enhanced wireless connectivity to meet the growing demand for intelligent, on-device processing in IoT applications.

"Definitely, AI is all impacting business models and cellular use cases — that ability to make decisions on the edge, dissecting data locally so that only the relevant information gets sent back, has really allowed for less overall data traffic. This, in turn, creates better business cost opportunities, easier cloud integration, and ultimately a more efficient system."
Jim Engleson, director of business development at Fibocom

Drivers for edge AI and IoT connectivity modules

Edge AI minimises data transfer and enhances security. Integrating AI at the edge enables intelligent computation at the data's origin, transmitting only ▶



Fibocom's mowing robot demo at Embedded World 2024

Fibocom's intelligent virtual boundary mowing robot

(Source: IoT Analytics)

Beyond operational efficiency, edge AI disrupts reliance on hyperscale cloud providers, making advanced AI more accessible and fostering decentralised innovation

actionable insights, such as metadata and diagnostics, instead of large datasets. This reduces network strain, optimises resources and enhances privacy by limiting sensitive data exposure. In industries like healthcare and manufacturing, it also mitigates risks from cloud outages or connectivity issues.

Edge AI reduces cloud reliance. Beyond operational efficiency, edge AI disrupts reliance on hyperscale cloud providers, making advanced AI more accessible and fostering decentralised innovation. This shift enables autonomous adaptability in systems like smart factories that self-optimise, vehicles reacting instantaneously to unpredictable situations in the surrounding environment and medical devices delivering AI-driven diagnostics at the point of care. As edge AI matures, it will redefine scalability, resilience and intelligence distribution globally.

Emerging use cases for edge-AI and IoT connectivity modules

Use case 1: Enhanced robotics

Robotics demand real-time AI with minimal cloud reliance. The demand for intelligent, autonomous systems in robotics is rising across industries. A key challenge is enabling real-time decision-making and efficient data processing while maintaining reliable connectivity and without heavy reliance on cloud computing, which can introduce latency, high transmission costs and privacy concerns.

Fibocom's SC171 module advances AI-driven autonomous systems. Fibocom is addressing this with its SC171 module, built on Qualcomm's QCM6490 chipset. The module integrates edge AI, 5G cellular IoT, and Wi-Fi 6E, providing a high-performance solution for autonomous devices. With its 8-core processor capable of delivering up to 12 TOPS of AI computing power, this module allows local data processing, reducing latency and improving efficiency. In practical applications, such as the Fibocom intelligent mowing robot, the SC171 module enables AI-driven lawn recognition and autonomous navigation. It supports real-time path planning, obstacle avoidance and precise edge-cutting, enhancing operational performance. Additionally, the robot's ability to autonomously return for recharging further boosts efficiency.

By deploying AI at the edge, the SC171 module minimizes cloud dependency, reduces data transmission costs, and enhances both privacy and operational efficiency, making it ideal for applications in industries like agriculture, industrial automation, and smart infrastructure. ▶

Use case 2: Deploying small language models at the micro/thin edge

Small language models (SLMs) boost real-time AI efficiency at the edge. SLMs are becoming increasingly efficient for edge deployments, enhancing real-time AI inference while reducing cloud reliance and eliminating associated costs. **DeepSeek**'s distilled R1 models have demonstrated how SLMs provide faster processing, lower latency and improved data privacy when deployed at the micro or thin edge.

Fibocom's AI Buddy enables real-time, cloud-free AI assistance. At CES 2025, Fibocom introduced AI Buddy, an edge AI-powered assistant designed to deliver real-time intelligence and seamless connectivity without relying on cloud processing. Powered by embedded versions of ChatGPT 4o and Claude 3.5 Sonnet, AI Buddy enables real-time natural language processing, context-aware responses and decision-making on the device itself. This autonomous functionality makes it ideal for hands-free, on-the-go applications.

AI Buddy's multifunctional design supports real-time translation, AI-driven image recognition and global data roaming, serving as a powerful tool for travelers and professionals. Optimized for hands-free interactions and seamless integration with smart wearables like earbuds, smart glasses and bracelets, AI Buddy delivers an intuitive, voice-controlled user experience — ushering in a new era of edge AI-driven personal assistants.

Looking ahead at cellular IoT and edge AI

IoT is evolving towards AI-driven automation and efficiency. The IoT industry is rapidly evolving from basic connectivity to intelligent, AI-driven automation, driven by advancements in 5G IoT, edge computing and AI integration. The convergence of these technologies is reshaping industries such as telecom, automotive, utilities and manufacturing, enabling real-time decision-making, resource optimization and reduced cloud dependency.

Market opportunities

The convergence of 5G IoT, edge computing and AI is opening more opportunities for vendors, adopters and investors alike.

Growth in cellular IoT and 5G expansion

- Cellular IoT connections are projected to grow at a 15% CAGR, reaching 9.7 billion by 2030.
- FWA powered by 5G is expected to grow at a 50% CAGR between 2024 and 2028, expanding high-speed connectivity to underserved regions.
- 5G RedCap is gaining traction as a cost-efficient, moderate-bandwidth IoT solution. By 2030, RedCap is projected to drive global cellular IoT connectivity, growing at an 18% CAGR from 2025 onward.
- The smart meter market is on a strong growth trajectory, with installations expected to reach 1.75 billion by 2030. India's US$20 billion investment in digital energy management is driving large-scale smart meter deployments, enhancing operational efficiency and resource optimisation.

Rise of AI-enabled cellular IoT and edge AI

- The shift towards edge AI is driving chipset vendors to integrate multi-core CPUs with NPUs/GPUs in their SoC designs.
- The adoption of AI-enabled cellular IoT is accelerating, with cellular connectivity and AI accelerators converging into single SoCs and IoT modules.
- AI-enabled cellular IoT modules are projected to grow at an 88% CAGR between 2024 and 2027, unlocking new use cases across industries. ▶

Key emerging trends

AI-enabled 5G modules in mobility

- AI-enabled cellular modules designed for mobility applications will see accelerated adoption, especially with 5G connectivity.
- By 2027, AI-enabled 5G modules for mobility applications are expected to make up 45% of all AI-enabled cellular module shipments.

AI in cellular LPWA: Unlocking new opportunities

While most cellular IoT modules currently focus on standard 5G and 4G technologies (with 2G and 3G as fallbacks), cellular low-power wide-area (LPWA) modules offer significant untapped potential. Integrating AI and TinyML into LPWA devices enables smarter, more responsive solutions across industries, allowing for edge AI-powered predictive maintenance, user behaviour analysis for automation and personalisation and new smart device categories that process data locally, reducing cloud dependency and improving efficiency.

Example: *A key innovation in this space is the Sony Altair ALT1350, a low-power LTE-M/NB-IoT SoC equipped with AI capabilities for low-power acceleration. Designed for edge processing and TinyML model inference, this chipset paves the way for AI-enabled cellular LPWA modules, allowing energy-efficient, AI-driven applications in sectors such as smart metering, asset tracking, and industrial applications.*

www.fibocom.com

Final thoughts

"As a global leading provider of AIoT communication modules and solutions, Fibocom embraces all industry advancements to deliver unparalleled value to our customers and partners worldwide. In the AI era, we will focus on AI-driven solutions that optimise cost, speed and efficiency through cloud, edge and edge-cloud integration. We believe these solutions will drive AI transformation across industries such as consumer electronics, FWA, lawnmowers and smart retail."
Eva Chen, vice president of Strategic Marketing at Fibocom

"AI integration in IoT modules reduces network data transmission, enabling autonomous decision-making and more compact designs of the thin and micro edge IoT devices. Building on smart modules, this shift drives a new product category – intelligent IoT modules – led by AI-enabled chipset providers like Qualcomm. As IoT moves beyond basic connectivity, IoT chipset players will be integrating high-performance multi-core CPUs, specialised GPUs and NPUs into their SoC/SiC designs to accelerate this transformation."
Satyajit Sinha, principal analyst at IoT Analytics

"There's no escaping that AI advances are opening up compelling new opportunities for driving operational efficiencies and new functionality within IoT smart modules. It's clear that greater intelligence at the IoT module level is enabling more to be done on the device, reducing the need for network communications, centralised data processing and potentially resulting in significantly lower environmental impacts from IoT thanks to power consumption optimisation."
George Malim, managing editor, IoT Now