Publication date:
29 Nov 2024

Author:
Kerem Arsal

QCT advances 5G edge with a new carrier-grade server and the QCT 5G x AI Dev. Kit

QCT partners with Intel to boost the adoption of 5G edge and AI

Brought to you by Informa Tech

Commissioned by:

Contents

Summary

As the industry shifts toward 5G standalone (SA) architectures and interest in artificial intelligence (AI) continues to grow, telecom operators worldwide are increasingly recognizing the significant opportunities presented by 5G edge. To seize these opportunities, operators are expanding their 5G edge infrastructures. This expansion is motivated by internal factors, such as reducing network costs and achieving operational efficiencies, as well as external factors, including improving customer experiences and creating new revenue streams.

Since 2020, Quanta Cloud Technology (QCT), in partnership with Intel, has been providing telecom operators and enterprises with carrier-grade servers. With its QuantaEdge line of servers powered by Intel® Xeon® Scalable processors, QCT offers operators flexible, compact computing infrastructures that are ideal for edge locations. These servers can run both virtualized 5G network functions and edge/mobile edge computing (MEC) applications.

This whitepaper discusses two products launched by QCT in 2024. The first product is the new QuantaEdge EGX77B-1U ultra short depth server, featuring 4th Gen Intel® Xeon® Scalable processors with Intel® vRAN Boost. This integrated solution eliminates the need for external hardware acceleration for vRAN workloads. The second product is the QCT 5G x AI Dev. Kit, designed to simplify the validation processes and implementations of AI applications for enterprise use cases, thereby lowering the barriers to AI adoption.

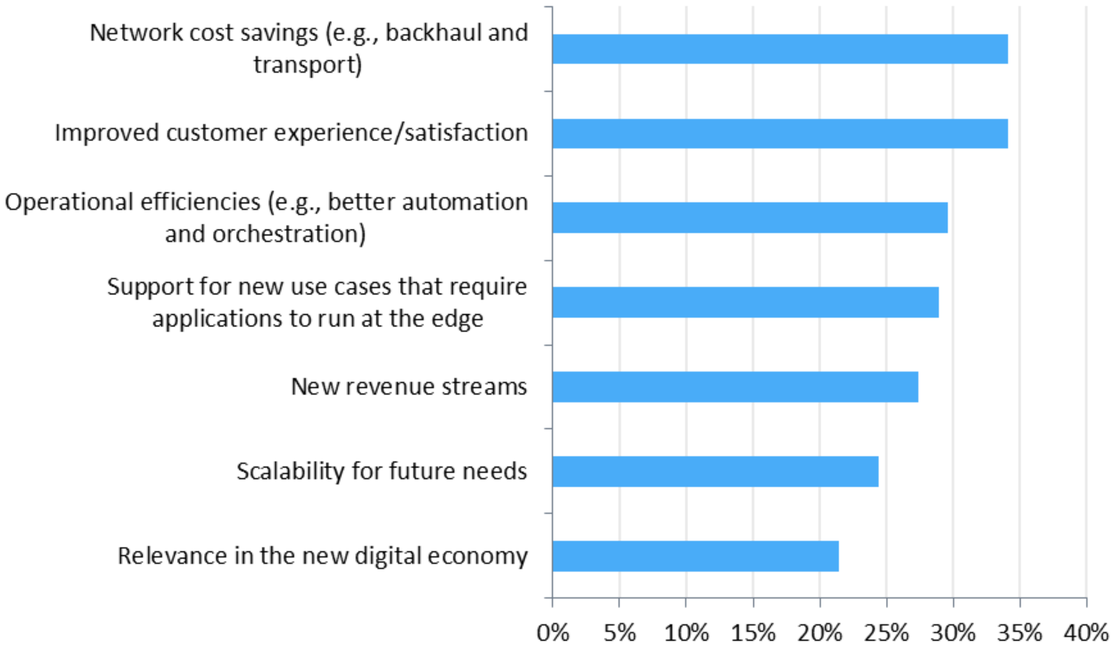
5G edge is a rising priority for operators

Telecom operators value the versatility of edge computing on 5G networks

Telecom operators are increasingly focusing on the potential benefits of edge computing on 5G networks, particularly as they transition to SA architectures. They view 5G edge as an extremely versatile proposition. By using shared hardware and software that can support both network functions and third-party IT workloads, operators aim to achieve positive outcomes both internally and externally.

According to Omdia's CSP Edge Computing 2024 survey, not only are operators expecting cost savings on their networks (34% of respondents) and better operational efficiencies (30% of respondents), but they are also equally motivated by external drivers, such as better customer experiences (34% of respondents) and new revenue streams (27% of respondents). The multi-faceted nature of edge computing makes it very powerful and appealing for telecom operators.

Figure 1: Incentives for communications service providers (CSPs) to deploy edge computing on their networks



Notes: n=135, Omdia CSP Edge Computing survey, fielded May–June 2024. Sample from the US, the UK, and Germany

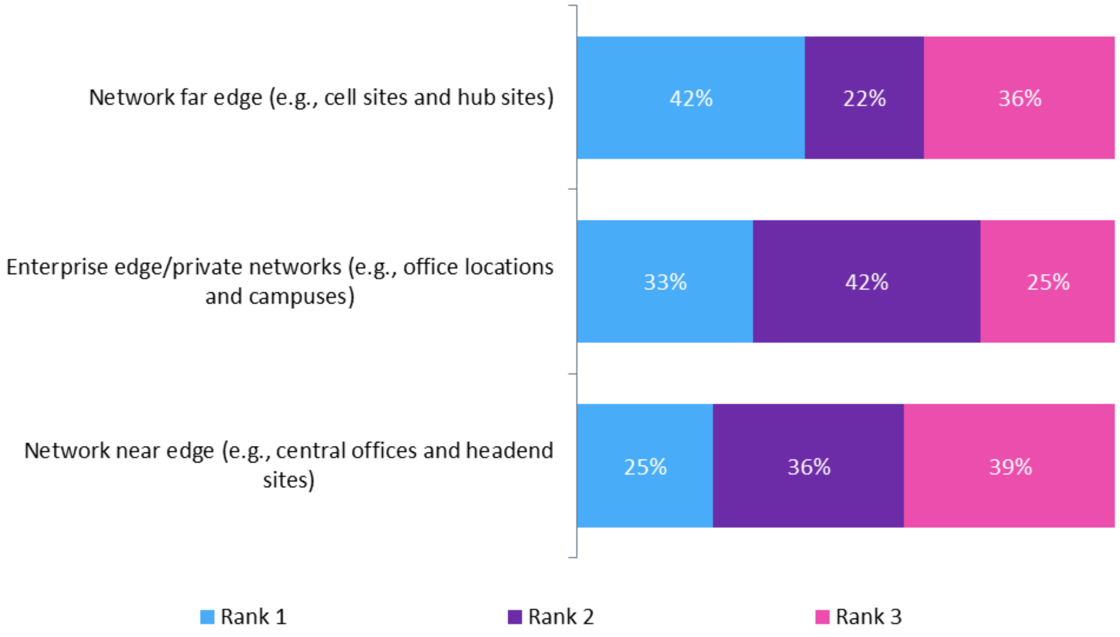
© 2024 Omdia

Source: Omdia

Operators are expanding their 5G edge infrastructures

Most early edge deployments in telco networks have occurred at centralized metro nodes. However, operators are now planning to shift toward more distributed computing topologies in their networks. According to Omdia's CSP Edge Computing 2024 survey, operators show significant interest in expanding their edge infrastructures to the far nodes of their networks, such as cell sites and small hub locations, in the next few years. In fact, 42% of CSP respondents in the US, the UK, and Germany ranked these locations as their top near-term priority for networked edge development.

Figure 2: CSP priorities for edge deployment locations over the next few years



Notes: n=135, Omdia CSP Edge Computing survey, fielded May–June 2024. Sample from the US, the UK, and Germany. Respondents were asked to rank their top 3 choices

© 2024 Omdia

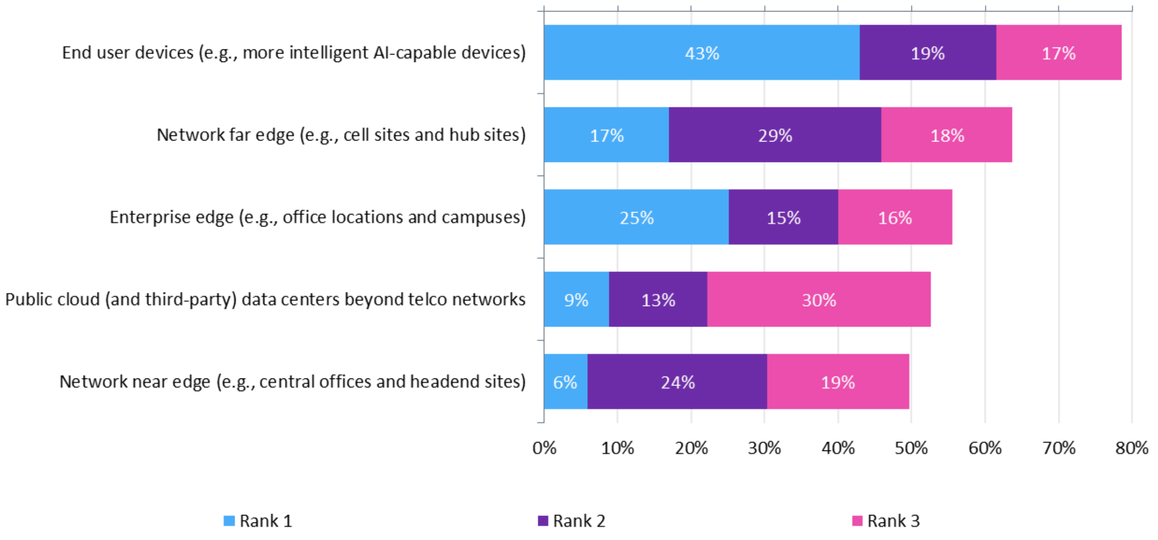
Source: Omdia

The far edges of 5G networks are becoming key computing locations for AI workloads

New artificial intelligence and machine learning (AI/ML) workloads will play a crucial role in expanding computing capabilities to the far edges of 5G networks, particularly at access points and aggregation nodes. While training large foundation models often occurs in very large data centers that host massive GPU clusters, inferencing—the actual execution of tasks using these models—and some small-scale model trainings are more likely to happen closer to end users or devices. This shift is driving the demand for AI computing capabilities in telecom networks.

As a result, telecom operators have a significant opportunity to host and run AI/ML processing at the far edges of their networks. In fact, 64% of CSP survey respondents agreed that far edge nodes would become the second most important locations for AI workloads, following end devices.

Figure 3: CSP expectations regarding where most of AI/ML processing will occur



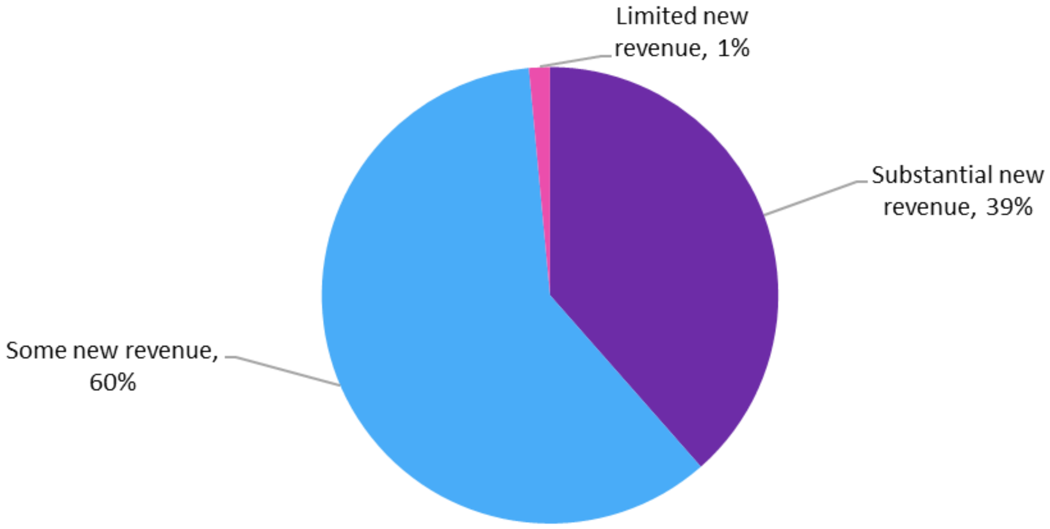
Notes: n=135, Omdia CSP Edge Computing survey, fielded May–June 2024. Sample from the US, the UK, and Germany. Respondents were asked to rank their top 3 choices

© 2024 Omdia

Source: Omdia

With these drivers in place, operators are turning to edge computing on their networks to solve some of their 5G monetization challenges. Considering the growing interest in AI, operators have renewed their enthusiasm and optimism toward potential revenue from 5G edge applications. According to Omdia's survey, a staggering 99% of operators in the US, the UK, and Germany reported that they are expecting to generate new revenue from edge applications, with 39% of them targeting substantial new revenue in this area.

Figure 4: Revenue expectations for CSPs from edge computing on their networks



Notes: n=135, Omdia CSP Edge Computing survey, fielded May–June 2024. Sample from the US, the UK, and Germany

© 2024 Omdia

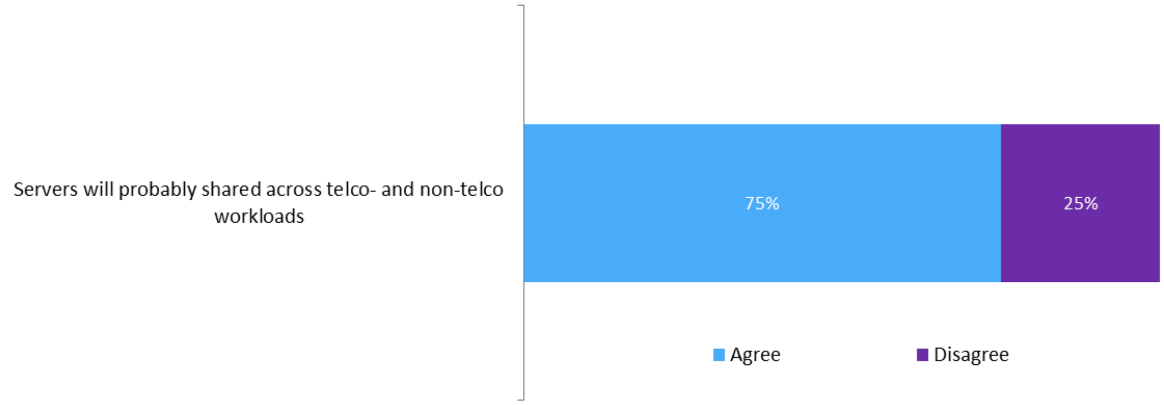
Source: Omdia

QCT’s QuantaEdge EGX77B-1U carrier-grade server is ideal for edge applications in 5G networks

QCT’s QuantaEdge EGX77B-1U server fits the requirements of edge computing in 5G networks perfectly

Flexibility across workloads without compromising RAN performance: Operators aiming to embrace 5G edge computing are keen to adopt hardware that can be shared across their network functions and IT workloads, especially for edge and MEC use cases. According to Omdia’s research, three out of four operators envision this future.

Figure 5: CSPs' long-term vision on using shared hardware to run telco workloads (e.g., network functions) and IT workloads (e.g., IT and edge applications) on telco cloud



Notes: n=135, Omdia CSP Edge Computing survey, fielded May–June 2024.
Sample from the US, the UK, and Germany

© 2024 Omdia

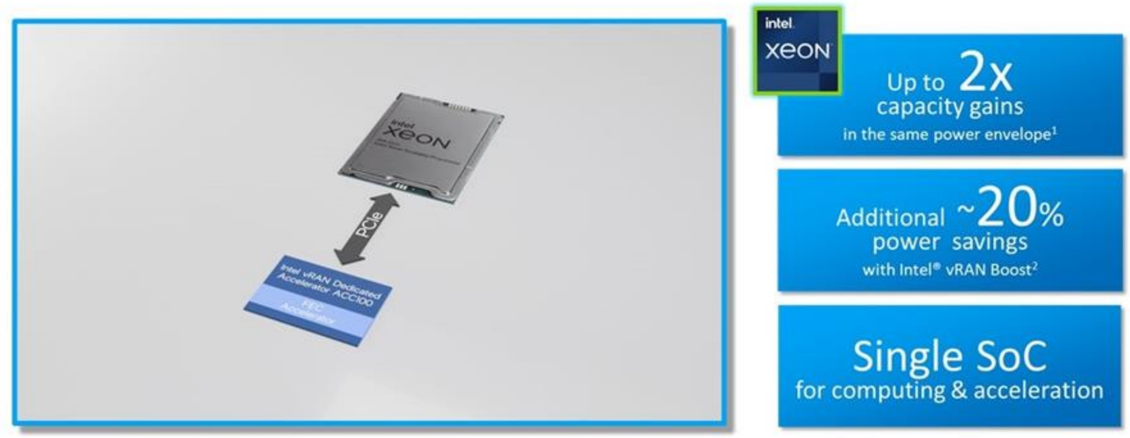
Source: Omdia

Most general-purpose servers offer such inherent flexibility, but purpose-built network equipment is commonly deployed separately from edge applications. Of course, it is important to note that flexibility should not compromise network performance because operators should prioritize maintaining high-quality computing for 5G connectivity. By using servers with integrated acceleration capabilities in their processors, they can achieve excellent performance while simplifying the overall system.

QCT's new carrier-grade edge server, the QuantaEdge EGX77B-1U, provides this combination of flexibility and performance. It is equipped with 4th Gen Intel® Xeon® Scalable processors with Intel® vRAN Boost. This server integrates a single system-on-chip (SoC) for both computing and acceleration, making it an ideal choice for running network and edge/MEC applications.

The Intel® vRAN Boost provides integrated acceleration for vRAN network functions, claiming up to a 2x capacity increase for vRAN and an additional 20% reduction in power consumption compared with the previous generation. This technology has already been deployed commercially by several Tier 1 operators around the world.

Figure 6: 4th Gen Intel® Xeon® Scalable processors with Intel® vRAN Boost in QCT's QuantaEdge EGX77B-1U carrier-grade server


intel
xeon
PCIe
Intel vRAN Dedicated Accelerator ACC100
FEC Accelerator
intel
xeon
Up to 2x
capacity gains
in the same power envelope[1]
Additional ~20%
power savings
with Intel® vRAN Boost[2]
Single SoC
for computing & acceleration

Notes: 1, 2= For specific workloads and configurations, see Intel® Performance Index.

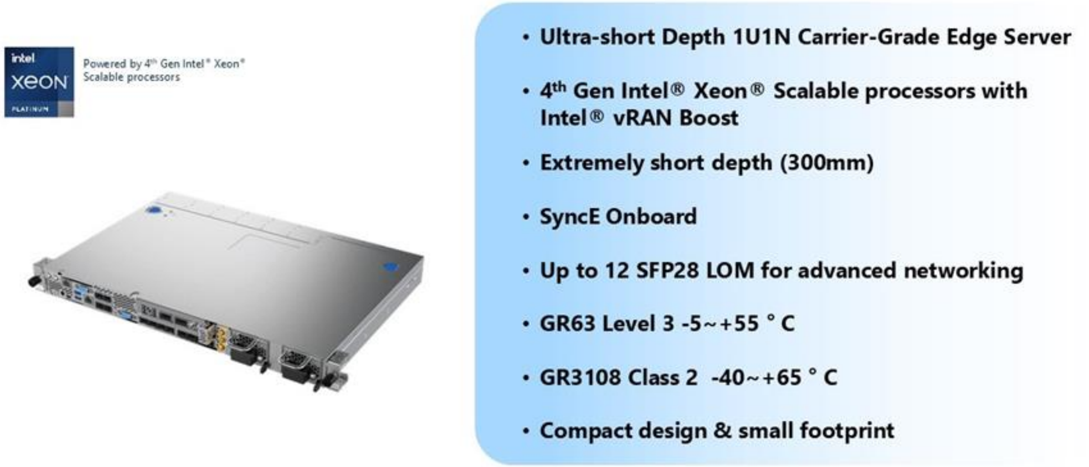
Source: Intel®

Compact design and wide operating temperature range for deployment at far edge nodes: Deploying additional edge computing hardware at far edge nodes, such as access and aggregation points, can pose serious challenges for telecom operators owing to limited space. Without sufficient space, operators must either forgo the 5G edge opportunities at these locations or pay significant costs to upgrade their sites to accommodate edge computing. To avoid these unwanted outcomes, it is critical for edge servers to be optimized in size and configured adequately for these settings.

Some edge nodes also present challenges related to operating conditions. Unlike the highly controlled data center environments, these locations can experience significant variables in temperature and humidity. If servers cannot function reliably under these conditions, operators may face downtime and computing inefficiencies, which can lead to significant maintenance costs and a potential need for replacements.

QCT's EGX77B-1U server is a great fit for such environments in terms of space and adaptability to varying operating conditions. With an extremely short depth of approximately 300mm, this edge server easily fits into racks and cell site cabinets. Its slim form factor allows operators to use their existing space more efficiently, potentially reducing costs associated with edge deployment at the far edges of 5G networks that otherwise require site upgrades. Additionally, QCT's EGX77B-1U server operates reliably between a temperature range of -40^{o}C to 65^{o}C, an operating range that should cover nearly all realistic edge locations.

Figure 7: QCT’s EGX77B-1U is an ultra-short depth carrier-grade server for 5G and MEC



- Ultra-short Depth 1U1N Carrier-Grade Edge Server
- 4th Gen Intel® Xeon® Scalable processors with Intel® vRAN Boost
- Extremely short depth (300mm)
- SyncE Onboard
- Up to 12 SFP28 LOM for advanced networking
- GR63 Level 3 -5~+55 ° C
- GR3108 Class 2 -40~+65 ° C
- Compact design & small footprint

Source: QCT

In short, QCT’s new EGX77B-1U server presents an excellent solution for operators deploying edge services in 5G networks. This carrier-grade edge server allows operators to run both 5G and edge/MEC applications on a single piece of hardware without compromising capacity and performance. Additionally, its slim, extremely short-depth form factor and ability to function across a wide range of temperatures make QCT’s EGX77B-1U server well-suited for the physical constraints found at the far edges of 5G networks, helping operators avoid unnecessary and costly site upgrades.

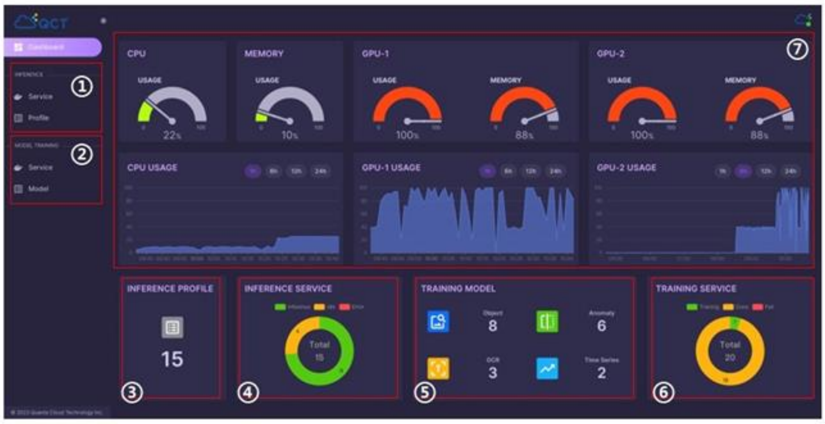
QCT 5G x AI Dev. Kit simplifies AI integration into 5G for enterprises

QCT 5G x AI Dev. Kit simplifies the validation of AI use cases for enterprises

The rate of AI adoption among enterprises will depend greatly on how easily enterprises can validate AI applications for their specific use cases before committing to larger-scale investments. If enterprises are unable to witness AI in action using their own data and usage contexts, they may struggle to set the right targets, become overwhelmed by technical complexities, and waste valuable time and money.

The QCT 5G x AI Dev. Kit supports the adoption of AI at the edge by significantly accelerating the validation of AI use cases for enterprises. With a simple, code-free GUI, the QCT 5G x AI Dev. Kit allows users to build and view the entire process of training and inference for AI models. It also allows for effective management of resource utilization for computing tasks.

Figure 8: The QCT 5G x AI Dev. Kit's GUI for AI training and inference processes



① Inference profile and service management

② Training model and service management

③ Number of inference profiles

④ Number and status of inference services

⑤ Types and number of inference models

- Default models: Object detection, OCR, anomaly detection, time-series prediction

⑥ Number and status of training services

⑦ System resource monitoring

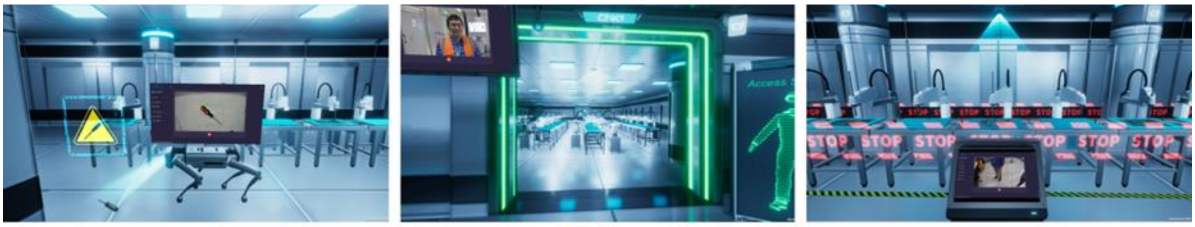
Source: QCT

To facilitate greater acceptance of AI adoption, the QCT 5G x AI Dev. Kit allows enterprises to experience AI for their own use cases quickly. Users can use their own private data to fine-tune existing models, such as those for object detection and optical character recognition, which are available in Intel® Distribution of OpenVINO™. By including several of the most used models in the QCT 5G x AI Dev. Kit by default, QCT enables enterprises to become familiar with real-life applications of AI technology.

For instance, object detection models are among the most versatile tools in the field of AI, as they can be adapted for various use cases. In manufacturing, object detection can be applied to robotics surveillance (e.g., conducting fault checks to ensure machines are functioning correctly); worker safety inspection (e.g., monitoring physical locations on the factory floor and ensuring the use of personal protective equipment); and virtual fencing (e.g., defining hazardous areas that trigger alerts if someone enters them).

It can also be used for predictive maintenance by monitoring the condition of equipment, for inventory and asset tracking by capturing and counting objects, and for quality assurance by constantly monitoring production processes or end products. For enterprises, having the ability to see a proof-of-concept for these applications quickly and to experience the tangible benefits of businesses firsthand can significantly reduce the challenges associated with adopting AI.

Figure 9: The object detection models in the QCT 5G x AI Dev. Kit enable many use cases in manufacturing



Robot Surveillance
Inspect the obstacles on the working area

AI Safety Inspection
Detect if workers wear safety gear

Virtual Fence
Detect if workers step into the danger zone

Source: QCT

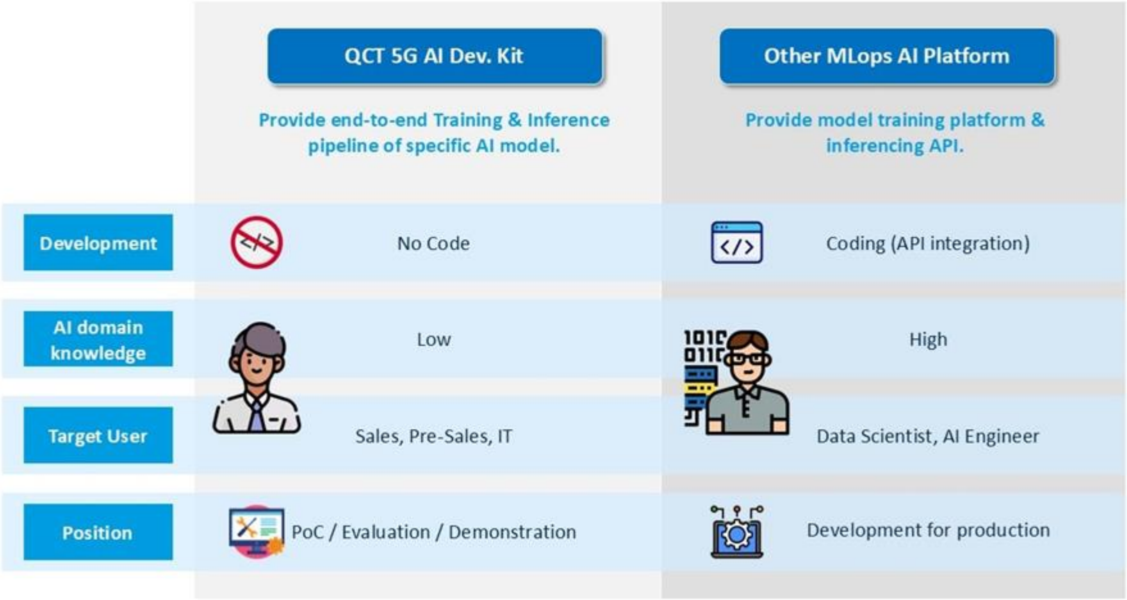
QCT 5G x AI Dev. Kit reduces barriers for enterprises

Accessing a development kit that simplifies the validation process of AI applications eliminates the need for costly upfront investments and overly complex deployments. For example:

- **Building an AI application internally from scratch** requires significant in-house AI skills and expertise, along with unpredictable spending on modelling, training, and application development.

- **Buying pre-developed solutions** often requires extensive customization and presents challenges in aligning various stakeholders, including system integrators, network providers, application developers, and digital infrastructure companies.

Figure 10: Market positioning of the QCT 5G x AI Dev. Kit vs. other approaches



	QCT 5G AI Dev. Kit	Other MLops AI Platform
	Provide end-to-end Training & Inference pipeline of specific AI model.	Provide model training platform & inferencing API.
Development	No Code	Coding (API integration)
AI domain knowledge	Low	High
Target User	Sales, Pre-Sales, IT	Data Scientist, AI Engineer
Position	PoC / Evaluation / Demonstration	Development for production

Source: QCT

In summary, the QCT 5G x AI Dev. Kit simplifies the integration of AI technology, making it more accessible for enterprises that have previously hesitated to adopt AI in their operations owing to concerns about complexity. This flexibility significantly lowers the barrier to entry for these businesses.

Appendix

Methodology

This report leverages primary research data from Omdia's CSP Edge Computing survey, fielded from May to June 2024, involving 135 telecom operator respondents in the US, the UK, and Germany. The survey was not commissioned by QCT or Intel®. For information and opinions on the current state and expected future of the 5G edge market, Omdia has relied on its own internal reports and expertise. For more specific product-related information and conclusions, this whitepaper uses documentation provided by QCT, which covers specifications about the EGX77B-1U server and the QCT 5G x AI Dev. Kit.

Author

Kerem Arsal
Senior Principal Analyst, Networked Edge
kerem.arsal@omdia.com

Get in touch

www.omdia.com
askananalyst@omdia.com

Omdia consulting

Omdia is a market-leading data, research, and consulting business focused on helping digital service providers, technology companies, and enterprise decision-makers thrive in the connected digital economy. Through our global base of analysts, we offer expert analysis and strategic insight across the IT, telecoms, and media industries.

We create business advantage for our customers by providing actionable insight to support business planning, product development, and go-to-market initiatives.

Our unique combination of authoritative data, market analysis, and vertical industry expertise is designed to empower decision-making, helping our clients profit from new technologies and capitalize on evolving business models.

Omdia is part of Informa Tech, a B2B information services business serving the technology, media, and telecoms sector. The Informa group is listed on the London Stock Exchange.

We hope that this analysis will help you make informed and imaginative business decisions. If you have further requirements, Omdia's consulting team may be able to help your company identify future trends and opportunities.

Copyright notice and disclaimer

The Omdia research, data and information referenced herein (the “Omdia Materials”) are the copyrighted property of Informa Tech and its subsidiaries or affiliates (together “Informa Tech”) or its third party data providers and represent data, research, opinions, or viewpoints published by Informa Tech, and are not representations of fact.

The Omdia Materials reflect information and opinions from the original publication date and not from the date of this document. The information and opinions expressed in the Omdia Materials are subject to change without notice and Informa Tech does not have any duty or responsibility to update the Omdia Materials or this publication as a result.

Omdia Materials are delivered on an “as-is” and “as-available” basis. No representation or warranty, express or implied, is made as to the fairness, accuracy, completeness, or correctness of the information, opinions, and conclusions contained in Omdia Materials.

To the maximum extent permitted by law, Informa Tech and its affiliates, officers, directors, employees, agents, and third party data providers disclaim any liability (including, without limitation, any liability arising from fault or negligence) as to the accuracy or completeness or use of the Omdia Materials. Informa Tech will not, under any circumstance whatsoever, be liable for any trading, investment, commercial, or other decisions based on or made in reliance of the Omdia Materials.