



QCT Enterprise 5G End-to-End Use Case Solution - SmartAMR

A Video Surveillance with Optimized Inference Performance on Validated E5G Network

Video Surveillance

Market Trend

- The video surveillance market is forecasted to reach 69.1 billion US dollars in 2026.
- Enterprise 5G optimizes network performance and lowers the latency, enabling high-quality video streaming and high-speed data transmission.
- QCT provides a comprehensive end-to-end solution for enterprises to enhance security monitoring by adopting an AI-driven video surveillance system.

Solution Key Highlights

- Combines AI with QCT E5G to enable massive uplink throughput and guaranteed low latency for real-time video analytics.
- Moves computing power to QCT's MEC servers for AI inference and workload reduction on end devices.
- Optimizes parameter configurations under QCT validated E5G to ensure a high-performance and reliable solution.

Executive Summary

The development of 5G technology is on the way to reach its maturity, bringing unlimited potentials for telco solution providers to differentiate their offerings and unleashing the limitation to fully demonstrate their capabilities. Featuring Enhanced Mobile Broadband (eMBB), Ultra Reliable Low Latency Communications (URLLC), and massive Machine Type Communication (mMTC), 5G technology enables high-quality video streaming and high-speed data transmission without any hindrance or data distortion. These are undoubtedly salient points that highly facilitates players devoted to developing or seeking video surveillance solutions. Benefited from innovative 5G communication technology, video surveillance will, to be sure, become a pivotal application for diverse industries on the market.

Seamlessly integrated with 5G network, video surveillance overcomes network challenges and provides high-applicable multimedia services with enhanced property security, gaining its appeal to more and more customers in the fields such as factories, farmlands, oil rigs, and many others. According to MarketsandMarkets¹, the high potential of video surveillance is expected to drive high demand and attract large investments, with the video surveillance market forecast to reach 69.1 billion US dollars in 2026, growing at a CAGR of 10.0% from 2021 to 2026.

As a Telco solution provider, QCT is ready to seize valuable opportunities with enterprises and launch an end-to-end use case solution with validated Enterprise 5G network (E5G). By moving AI computing power to the edge site, this redefined architecture can highly reduce end-device workloads and implement real-time applications under E5G network. To help facilitate enterprises boost their operational efficiency and reduce labor costs, QCT is devoted to optimizing the inference performance and lowering the latency for video streaming by offering its comprehensive and robust ecosystem of hardware, technology, and know-how experience.

Solution Overview

SmartAMR is a solution that adopts an Autonomous Mobile Robot (AMR) with AI analytics not only used to autonomously move materials without human intervention but also serve as a video surveillance system for security events such as object tracking, intrusion detection, and face recognition. Based on the image data obtained from camera resources, AI-driven analytics integrated on a high-mobile AMR simplifies the decision-making process and precisely analyzes the data, allowing enterprises to optimize workflow and improve productivity. To ensure high-reliability network performance, QCT E5G is validated with comprehensive end-to-end tests, providing massive uplink bandwidth and guaranteeing low latency required for real-time streaming analytics.

With enabled AI and QCT E5G on an AMR, QCT redefined its solution architecture by placing AI computing at the edge server, resolving insufficient computing capabilities at end devices. AI inference on edge server under E5G network is capable of streaming high-resolution images for detecting diverse objects, unexpected intruder, and face authentication in real time. With merely an application installed on any mobile or wearable device, enterprises can enjoy high-quality and real-time streaming experiences based on edge AI analysis for different scenarios.

¹ Market Trend for Video Surveillance: <https://www.marketsandmarkets.com/Market-Reports/video-surveillance-market-645.html>

Key Technology

eMBB

Defined by 3rd Generation Partnership Project (3GPP), 5G network provides three classes of services – enhanced Mobile Broadband (eMBB), ultra-Reliable Low Latency Communications (uRLLC), and massive Machine Type Communications (mMTC). eMBB has been rolled out to be the initial phase of 5G deployment for the real-world applications by adopting two models – Non-Standalone (NSA) and Standalone (SA) architectures. Being capable of delivering a data rate of 1 Gbps and beyond, eMBB delivers extraordinarily high bandwidth and low latency for implementing applications such as 4K media, augmented reality (AR), and virtual reality (VR).

Frame Structure Configuration

5G NR frame structure defined in 3GPP TS 38.211 specifies a single frame consists of 10 subframes. The duration of each subframe is 1 millisecond and each subframe is composed of 2 ^{μ} slots. Each slot can be configured fully utilized for Uplink (U), Downlink (D), or Special (S) (i.e. mixed uplink and downlink) transmissions. Thus, a single frame is composed of D, U, and S slots that are configurable based on the requirements of DL and UL bandwidth. Several common configurations of frame structure are DDDSUDDSUU, DDSUU, DSUUU, DDDSU, and DDDDDDDSUU, for example.

Inference

Inference is a technology adopted to detect objects, identify instances, and frame objects from images. Inference provides semantic understanding of images and videos for a variety of applications such as human behavior analysis, facial recognition, and autonomous driving. An inference model is generally trained by a convolutional neural network (CNN) to detect objects for different scenarios. With the inference model, the detected object can be recognized by combining multiple low-level image features into high-level context. Some common CNN models include Yolov4, Faster R-CNN, and Mask R-CNN.

OpenVINO™

Intel® Distribution of OpenVINO™ is a comprehensive toolkit used to quickly develop applications and solutions to solve a variety of tasks, including emulation of human vision, automatic speech recognition, natural language processing, recommendation systems, and many others. This toolkit enables Convolutional Neural Networks (CNN)-based deep learning inference on the edge and extends computer vision and non-vision workloads across Intel® hardware to optimize inference performance.

Hardware Configuration

Enterprise 5G Configuration

QCT provides a comprehensive Enterprise 5G solution with Enterprise 5G network, consisting of 5G Core Network (5GC) and 5G Radio Access Network (5G RAN). On the hardware layer, QuantaGrid D52BQ-2U with high density and high performance is best-in-class for building 5G Core (5GC) network, as shown in Table 1. Quanta Grid D52Y-2U with merely 400mm depth is best-suited for building 5G RAN, which can be closely located to end users. This server can support up to five PCIe expansion slots in front-access design for FPGA accelerator cards and high-speed network cards, allowing operators to easily replace cards from the front panel. In addition, the server that integrates Centralized Unit (CU) and Distributed Unit (DU) functions in one box can be adopted as a Baseband Unit (BBU) server, as shown in Table 2.

Table 1. Specification of 5G Core Server - QuantaGrid D52BQ-2U.




Product: 5G Core Server Model Name: QuantaGrid D52BQ-2U Dimension: 440 x 87.5 x 780 (mm)		
Items	Description	Amounts
Processor	Intel® Xeon® Processor Scalable Family	2
Memory	Up to 7.5TB (512Gx12+128Gx12) of memory for RDIMM/LRDIMM	16
Storage	Front Storage: (12) 3.5"/2.5" hot-plug SATA/SAS Rear Storage: (2) 2.5" hot-plug NVMe/SATA/SAS (optional)	14
Expansion Slots	(1) PCIe Gen3 x16 SAS mezzanine slot (1) PCIe Gen3 x16 OCP 2.0 mezzanine slot or PHY card (2) PCIe Gen3 x 8 FHHL or (1) PCIe Gen3 x16 FHHL (3) PCIe Gen3 x 8 FHHL or (1) PCIe Gen3 x16 + x8 FHHL (1) PCIe Gen3 x16 LP MD-2	Up to 8
Power Supply	1+1 High efficiency redundant hot-pluggable 800W/1200W PSU	2

Table 2. Specification of 5G RAN Baseband Unit - QuantaGrid D52Y-2U.

Product: Baseband Unit Model Name: OmniRAN-E5GBBU Dimension: 447.8 x 86.3 x 420 (mm)		
Items	Description	Amounts
Processor	Intel® Xeon® Processor Scalable Family	2
Memory	Up to 2TB (128Gx16) of memory for RDIMM/LRDIMM	16
Storage	(2) 2.5" hot-plug drives	2
Expansion Slot	(2) FHFL PCIe Gen3x16 (1) FHHL PCIe Gen3x16 (1) HHHH PCIe Gen3x16 or (2) HHHH PCIe Gen3x8 or (1) SAS Mezz adapter	Up to 5
Power Supply	1+1 High-efficiency redundant hot-pluggable 1600W PSU	2


Adopted as a Fronthaul Gateway (FHGW), IronRAN-FG GenA acts as an expansion unit for a Baseband Unit (BBU) server, as shown in Table 3. FHGW distributes downlink signals from a BBU server to multiple RRUs and merges uplink signals from multiple RRUs to a BBU server, thereby extending the coverage of 5G radio signals across a wide area. In addition, IronRAN-FG GenA with embedded GPS module is capable of connecting GPS antenna and sending IEEE 1588 timing packets respectively to BBU and RRU for time synchronization.

Table 3. Specification of 5G RAN Fronthaul Gateway - IronRAN-FG GenA.

Product: Fronthaul Gateway Model Name: IronRAN-FG GenA Dimension: 251.2 x 43.8 x 447.8 (mm)		
Items	Description	
Max. Number of Cell to Support	1 cell with 100MHz, 4x4 MIMO	
Max. Number of RRU to Support	4 RRUs	
Synchronization	IEEE 1588v2 (GPS module embedded)	
Optical Interface	SFP+ 10Gbps x 4 for RRU connection SFP+ 10Gbps x 1 for BBU connection Up to 2km for eCPRI data transmission by single mode fiber	
Power Consumption	< 70W without RRU < 400W with 4 x RRUs	

Adopted as a Remote Radio Unit (RRU), *IronRAN-RU1* PI GenA RRU is connected to a FHGW to convert digital signals and emit 5GNR wireless signals to 5G end devices, supporting the connectivity between 5G network and user equipment, as shown in Table 4. In addition, *IronRAN-RU1* PI GenA RRU supports O-RAN 7-2 Split in 5G RAN to mitigate the complexity in RRU and efficiently transport utilization by dividing PHY into high-PHY layer residing in DU and low-PHY layer residing in RU.

Table 4. Specification of 5G RAN Remote Radio Unit - IronRAN-RU1 PI GenA RRU

Product: Remote Radio Unit Model Name: IronRAN-RU1 PI GenA RRU Dimension: 204.7 x 259.6 x 52.5 (mm)		
Items	Description	
Frequency Band	n79: 4.8~4.9 GHz n78: 3.3~3.8 GHz	
Function Split	O-RAN option 7-2x category A	
Tx Power	250mW/channel, total: 1W (4T4R)	
Ingress Protection	IP30	
Power Consumption	< 60W	

Use Case Configuration

Quanta Grid D52Y-2U with ultra-short design is also recommended for building MEC servers, which can be closely located to end users and reduce the latency of data transmission (see Table 2). To ensure high-quality inference services at the edge sites, this server is equipped with powerful processors to perform timely data transmission and processing. Additionally, the server can support PCIe expansion slots for GPU cards, enhancing the capability of image streaming for inference calculation.

Solution Architecture

The end-to-end solution architecture is composed of an Autonomous Mobile Robot (AMR), 5G network, and AI compute infrastructure, as shown in Figure 1. Cameras embedded in the AMR are utilized to capture high-quality images by navigating the surrounding environment for video surveillance. Through 5G network, the captured image streams can be inferred at the MEC servers rather than at the local AMR.

Enterprise 5G network is validated with end-to-end tests to provide 5G Core (5GC) service and 5G RAN service. The end-to-end validation encompassing system performance, stability, and reliability ensures 5G system comprehensively meets commercial-level quality. Considering the continuous operation and failure mitigation for a network function, a network port, or an active server, 5GC is composed of two servers and two data switches for enabling high-availability (HA) mechanism. The failover process can be immediately activated to transfer services from an active server to a standby server. Further, 5G RAN consists of a BBU server, a Fronthaul Gateway (FHGW), and RRUs. The BBU server is running the DU and CU software, providing the baseband function with O-RAN Split Option 7-2. The FH-GW supports up to 4 RRUs for increasing Radio Coverage in the field. 5G RAN supports three frame structure configurations – DDDSUDDSUU, DDSUU, and DSUUU, satisfying the DL and UL bandwidth requirements for different types of applications.

In the AI computing infrastructure, three MEC servers are adopted to build a cluster. To enhance inference performance, OpenVINO™, a novel technology featuring layer fusion and precision calibration, is adopted to optimize facial recognition model. The recognition results can be transmitted to a centralized control center for real-time monitoring and auditing.

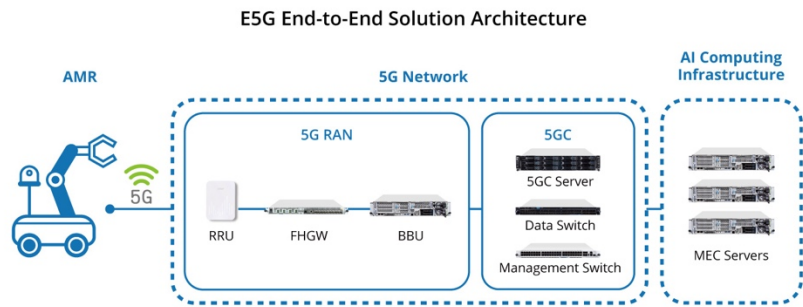


Figure 1. Solution Architecture.

Solution Performance

Based on the proposed architecture, inference throughput and end-to-end Round Trip Time (RTT) are the two test items conducted to validate the solution performance by configuring CPU plugin parameters in OpenVINO™. Testing inference throughput aims to figure out the maximum number of inference requests in a single MEC server. While, testing end-to-end RTT aims to estimate the duration, including 5G RTT and the latency of inference pipeline.

Test Result of Inference Throughput

Figure 2 illustrates the results of inference throughput under different video resolutions, including 640x480 pixels, 1280x720 pixels, and 1920x1080 pixels. After a series of tests for exploring optimal configurations in OpenVINO™, the test results generally reveal a sharp increase in the inference throughput in terms of the three video resolutions, indicating the inference engine can even process more frames in a second by optimizing CPU plugin parameters. However, as the video resolution increases, the performance of inference throughput slightly grows, which indicates the higher the video resolution is, the more the computing power is required for inference calculation. Overall, both performances with its respective parameter configured are capable of processing the number of frames two times more than those of in baseline performances. It is noted that by simultaneously configuring the two parameters, the performance can even achieve up to 3.8 times better than the baseline performance.

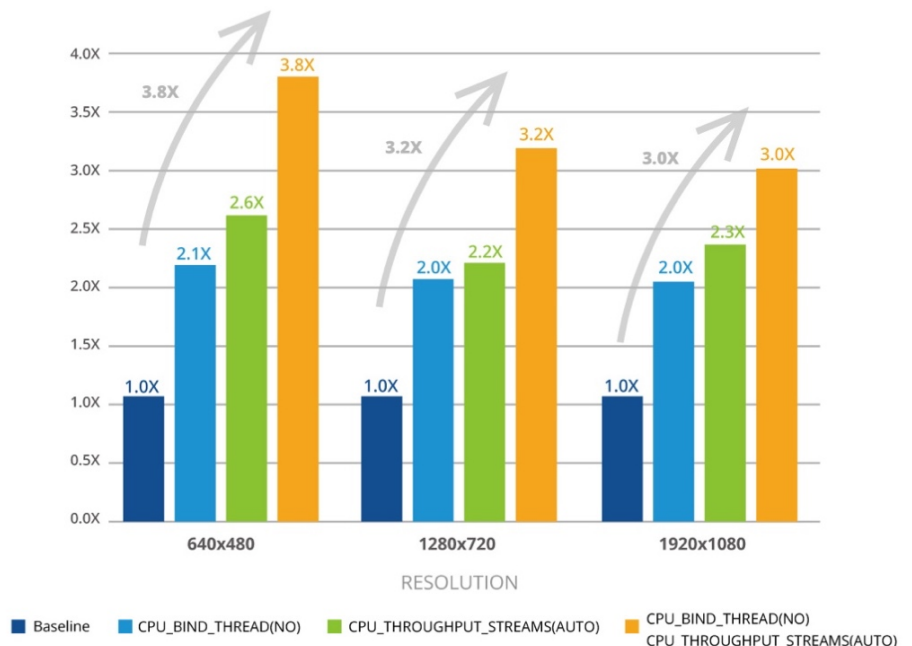


Figure 2. Performance of Inference Throughput.

Test Result of End-to-End RTT

Figure 3 illustrates the results of end-to-end RTT under different video resolutions, including 640x480 pixels, 1280x720 pixels, and 1920x1080 pixels. After a series of tests for exploring optimal configurations in OpenVINO™, the test results reveal a gradual decrease in RTT for different resolutions. As the video resolution increases, end-to-end RTT relatively climbs. This is due to the fact that the higher the video resolution is, the more the processing time for data transmission and consumption time are required for inference calculation. By adjusting the parameter configurations, even though the RTT generally grows as the resolution increases, the RTT declines in each individual resolution, ranging from 8 to 12% for each frame, compared to the baseline RTT.

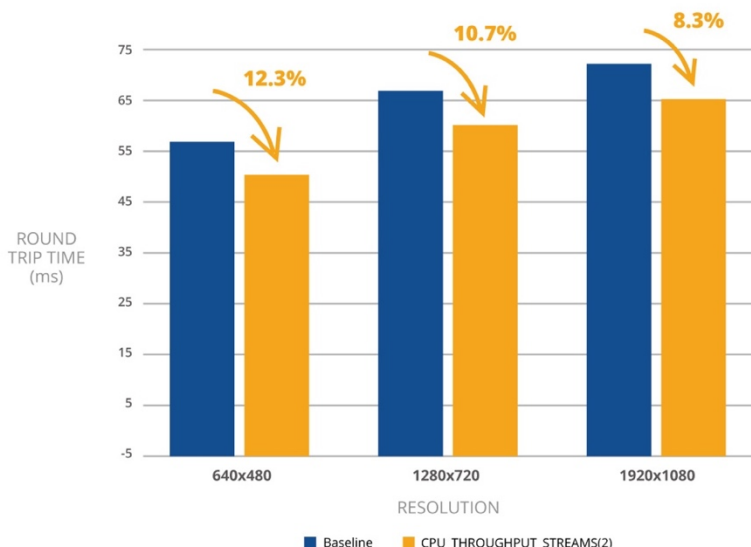


Figure 3. Performance of End-to-End Round Trip Time (RTT).

Conclusion

Driven by huge demand from digitalization transformation, 5G is set to break new ground to inevitably generate tremendous values on camera-based analytical applications such as intruder detection and intelligent image processing, allowing enterprises to introduce an all-round 5G technology with value-added safety and security functions. SmartAMR with real-time AI analysis on AMR takes the advantages of 5G to offer a robust and reliable video surveillance solution. This solution successfully demonstrates the capabilities to optimize throughput performance up to 3.8 times and significantly reduce latency up to 12% by adjusting parameters in OpenVINO, compared with the baseline performances. As a leading telco 5G infrastructure provider, QCT is well-prepared to take advantage of the booming video surveillance industry by providing the comprehensive hardware portfolio, AI know-how, and Enterprise 5G Solution.

ABOUT QCT

Quanta Cloud Technology (QCT) is a global data center solution provider. We combine the efficiency of hyperscale hardware with infrastructure software from a diversity of industry leaders to solve next-generation data center design and operation challenges. QCT serves cloud service providers, telecoms, and enterprises running public, hybrid and private clouds.

Product lines include hyperconverged and software-defined data center solutions as well as servers, storage, switches and integrated racks with a diverse ecosystem of hardware components and software partners. QCT designs, manufactures, integrates and services cutting-edge offerings via its own global network. The parent of QCT is Quanta Computer, Inc., a Fortune Global 500 corporation. <http://www.QCT.io>

United States

QCT LLC., Silicon Valley office
1010 Rincon Circle, San Jose, CA 95131
TOLL-FREE: 1-855-QCT-MUST
TEL: +1-510-270-6111
FAX: +1-510-270-6161
Support: +1-510-270-6216

China

云达科技,北京办公室 (Quanta Cloud Technology)
北京市朝阳区东大桥路 12 号润诚中心 2 号楼
TEL: +86-10-5920-7600
FAX: +86-10-5981-7958

云达科技,杭州办公室 (Quanta Cloud Technology)

浙江省杭州市西湖区古墩路浙商财富中心 4 号楼 303 室
TEL: +86-571-2819-8650

Japan

Quanta Cloud Technology Japan 株式会社
日本国東京都港区芝大門二丁目五番八号
牧田ビル 3 階
TEL: +81-3-5777-0818
FAX: +81-3-5777-0819

Germany

Quanta Cloud Technology Germany GmbH
Hamborner Str. 55, 40472 Düsseldorf
TEL: +49-2405-4083-1300

Korea

QCT Korea, Inc. (주식회사)
큐씨티코리아
서울특별시 영등포구 의사당대로 97
교보증권빌딩 10 층, 07327
TEL: +82-10-5397-1412
FAX: +82-2-6336-6710

Other regions

Quanta Cloud Technology
No. 211 Wenhua 2nd Rd., Guishan Dist.,
Taoyuan City 33377, Taiwan
TEL: +886-3-327-2345
FAX: +886-3-397-4770