Unlocking the Potential of 5G Services with A5G Networks 5G Core and QCT Servers Accelerated by NVIDIA GH200 Grace Hopper Superchip

Introduction

The evolution of 5G core networks marks a significant leap from traditional mobile architectures, driven by the need to support ultra-low latency, massive device connectivity, and high-speed data transmission for a wide range of applications—from enhanced mobile broadband to mission-critical IoT. As 5G services scale globally, operators are increasingly challenged to meet demanding performance, scalability, and efficiency requirements, all while managing operational and capital expenditures. One emerging use case is robotics, where reliable connectivity is essential—this is of great importance whether robots are navigating a factory floor or transitioning between indoor and outdoor environments where handovers are required. In such scenarios, edge solutions not only ensure consistent 5G coverage, but also enable real-time AI inferencing and workload processing at the edge, supporting intelligent and autonomous operations.

To address these challenges, QCT introduces servers accelerated by the NVIDIA GH200 Grace Hopper™ Superchip, a cutting-edge infrastructure solution accelerated by combining NVIDIA Grace™ CPUs, NVIDIA Hopper GPUs and NVIDIA BlueField® DPUs. This powerful combination delivers exceptional computing performance, intelligent acceleration, and energy-efficient operation, making it ideally suited for the high demands of 5G services.

A key part of this solution is the A5G Networks 4G/5G Core software stack, which includes advanced components such as the A5G UPF (User Plane Function). The A5G platform is designed with a cloud-native, microservices-based architecture that supports AI/ML-driven traffic management, dynamic network slicing, and flexible deployment options. Running on QCT servers accelerated by the NVIDIA GH200 Grace Hopper Superchip, A5G Networks 4G/5G Core software stack enables operators to build high-performance, agile, and intelligent 5G core networks that can scale efficiently with user and application demands. A5G Networks flexible and efficient architecture and various network functions such as UPF can run on CPUs and/or DPUs with AI/ML components utilizing GPUs. This flexibility allows 5G control plane network functions and other application workloads to run on CPUs and reduces one's overall energy footprint.

This solution brief highlights the advantages operators and enterprises can realize by deploying A5G Networks 5G core network functions on QCT servers accelerated by the NVIDIA GH200 Grace Hopper Superchip. It outlines how the platform enhances network performance, reduces costs, and enables intelligent, flexible, and scalable network operations for the 5G era.

QCT QuantaGrid S74G-2U

The QuantaGrid S74G-2U is a state-of-the-art server powered by NVIDIA's pioneering GH200 Grace Hopper Superchip, delivering exceptional performance across AI, HPC and 5G telecom workloads. At its core is a tightly coupled architecture that combines a 72-core Arm-based

Grace CPU and the Hopper GPU, interconnected via NVIDIA NVLink™-C2C. This design facilitates a unified memory architecture, allowing seamless data sharing between CPU and GPU, which is particularly advantageous for memory-intensive tasks such as AI inference and high-performance computing (HPC), scientific computing, and rea-time 5G processing applications.

Feature	Spec
Processor	(1)NVIDIA Grace Hopper™ Superchip Max TDP 1000W NVIDIA NVLink™-C2C 900GB/s
Memory	Up to 480GB LPDDRX embedded 96GB / 144GB HBM3 GPU memory
Storage	(4) E1.S NVMe SSDs
Onboard Storage	(2) PCIe gen5 M.2 22110/2280
Expansion Slot	(3) PCIe 5.0 x16 FHFL dual width slots
Fans	(5) 6056 dual rotor fans (N+1 redundant)
PSU	1+1 High efficiency hot-pluggable 2000W PSUs, 80 plus titanium
Form Factor	2U
Dimensions	438 x 87.5 x 900mm (w x h x d) 17.24" x 3.44" x 35.43" (w x h x d)

The server is built upon NVIDIA's MGX™ architecture, a modular framework that offers flexibility and scalability. This architecture enables rapid adaptation to evolving technological requirements by supporting various configurations of CPUs, GPUs, and DPUs, ensuring futureproofing of the infrastructure.

In terms of memory, the QuantaGrid S74G-2U is equipped with 480 GB of LPDDR5X memory alongside 144GB of HBM3 memory within the GPU. This substantial memory capacity, combined with the high-bandwidth NVLink™-C2C interconnect, ensures efficient handling of large datasets and complex computations.

NVIDIA MGX

Bring Accelerated Computing to AI and HPC users with modular design

GPU   CPU   DPU

A single MGX architecture enables

QuantaGrid S74G-2U

AI   HPC+Data Analytics   Digital Twins   Cloud Services   Cloud Gaming   5G

Specifically for 5G use cases, the QuantaGrid S74G-2U is optimized for Core and RAN use cases including UPF offloading vRAN acceleration, and AI-driven RAN. The platform's support for PCIe Gen5 with BlueField® DPU, GPUDirect® RDMA, and high-speed Network Interface Cards (NICs) ensures low-latency, high-throughput data transfer—crucial for telecom
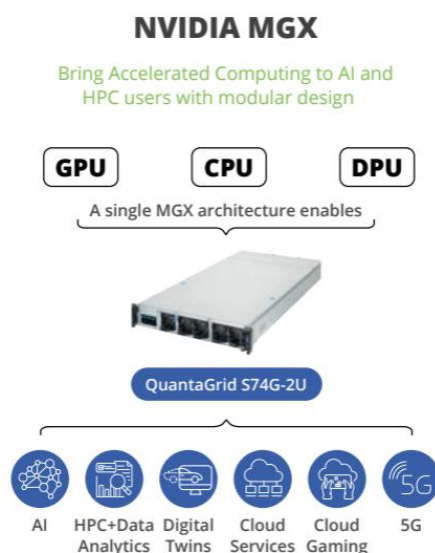
environments. With Arm System Ready certification, it ensures broad OS compatibility and seamless deployment into both cloud-native and traditional network infrastructures.

Overall, the QuantaGrid S74G-2U stands out as a robust and versatile platform that bridges compute-intensive AI/HPC demands with real-time, deterministic 5G network applications, enabling operators and enterprise to converge workloads on a single, future-ready infrastructure.

A5G Networks Product Highlight

The vision of A5G Networks is to enable and catalyze the upcoming transition to a distributed and autonomous mobile network for 4G, 5G and beyond. Its unique IP helps realize significant savings in capital and operating expenditure, reduces energy requirements, improves the quality of user experience and accelerates the adoption of new business models.

A5G Networks enables the deployment of enterprise private networks, connected car networks and distributed public mobile networks, while playing a pivotal role in a range of smart city projects.

The A5G Networks software uses a fully cloud-native, microservices-based architecture that scales elastically across hybrid and multi-cloud deployments.

A5G Networks packet core is a 3GPP R17-compliant 4G/5G converged core. The A5G UPF software is truly distributed and elastically scalable to meet the growing demand of user
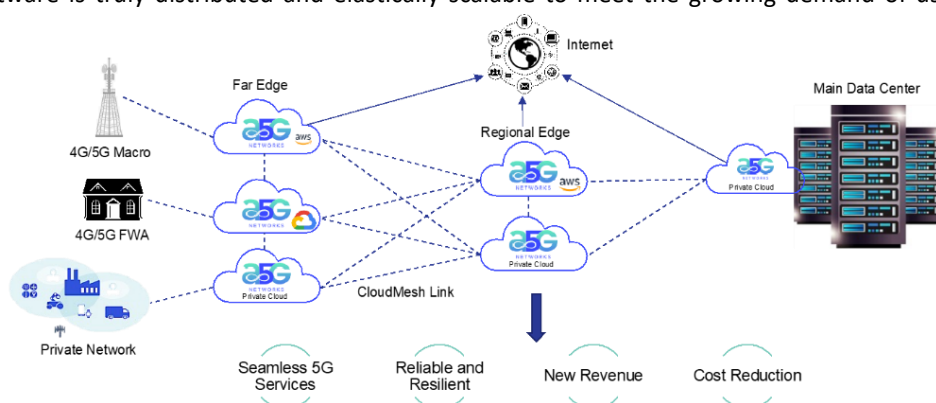
plane traffic and processing. It can support a wide variety of NIC cards, from non-DPDK-compatible traditional NICs to performance-centric DPDK-compatible SmartNICs. The A5G UPF software can provide high throughput by adding cores and high-capacity NICs, catering to many different use cases including Enhanced Mobile Broadband (eMBB), Ultra-Reliable Low-Latency Communications (uRLLC), and Massive Machine-Type Communications (mMTC), while providing ML-based optimizations to ensure that private and distributed networks are easy to manage and operate. The A5G UPF and SMF software can scale in a cluster deployment and provide throughputs in the order of 1Tbps. By leveraging NVIDIA and QCT technologies, the A5G Networks UPF software minimizes compute and energy requirements, leading to significant reductions in overall TCO.

Key Benefits and Features

In a next-generation 5G network architecture, performance, energy efficiency, intelligent traffic management, and deployment flexibility are critical success factors. QCT's QuantaGrid S74G-2U server, accelerated by the NVIDIA Grace™ CPU and NVIDIA BlueField® DPU, offers a

comprehensive solution that outperforms traditional pure software-based architectures across these dimensions.

1. Enhanced Performance and High Throughput

The QuantaGrid S74G-2U enables full deployment of the A5G 4G/5G core software stack on the NVIDIA Grace™ CPU, delivering exceptional computing performance. By offloading the UPF data path to the NVIDIA BlueField® DPU, the system leverages hardware acceleration to achieve near line-rate throughput—approximately 196 Gbps. Even with Deep Packet Inspection (DPI) enabled, the impact on UPF performance remains significantly less as compared to CPU based DPI. This setup delivers a substantial performance boost compared to pure software implementations, ensuring seamless support for high-bandwidth 5G services.

2. Significant Energy Efficiency and Cost Savings

Offloading UPF data plane traffic to the NVIDIA BlueField® DPU significantly reduces the host CPU workload and overall power consumption. Compared to traditional x86-based architectures, the NVIDIA Grace CPU delivers more than 50% power savings. When combined with the DPU's efficient processing capabilities, the overall system achieves over 50% power reduction compared to x86 CPUs, lowering both CAPEX and OPEX while supporting the development of green data centers.

UPF data plane flow were offloaded to NVIDIA BlueField® DPU instead of consuming host server CPU resources.
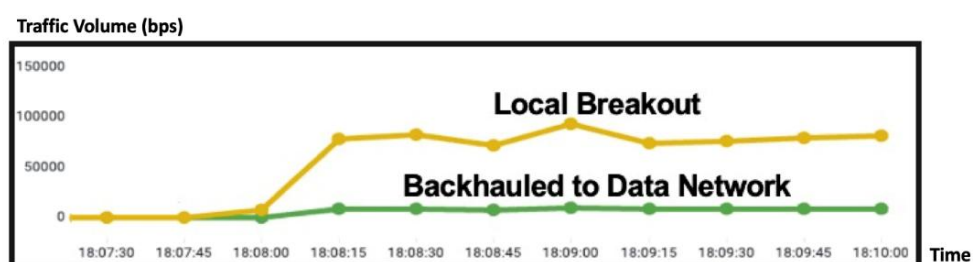
The figure shows the traffic behavior over time, comparing Local Breakout versus Backhauled to Data Network. The graph highlights that local breakout traffic (yellow) significantly outpaces backhauled traffic (green), showing the efficiency gained through traffic offloading.
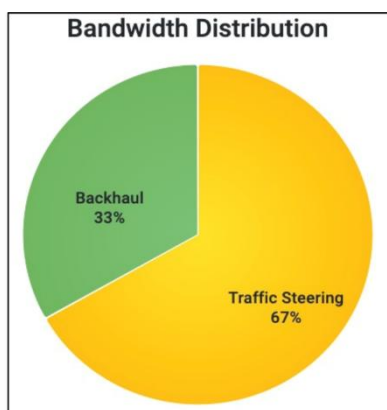
The figure illustrates the bandwidth distribution, where Traffic Steering (Local Breakout) accounts for 67% of the total bandwidth, while Backhaul uses only 33%, reflecting substantial reduction in backhaul dependency.

In summary, the solution sustains bi-directional 2x 100G line rate traffic while achieving more than 50% reduction in power consumption compared to traditional x86-based implementations. Additionally, there is a 67% reduction in backhaul traffic, significantly alleviating core network load and improving overall system efficiency.

3. Intelligent Traffic Management and Quality of Service

The A5G UPF integrates AI/ML-powered Deep Packet Inspection (DPI) and traffic classification to intelligently differentiate between real-time and non-real-time applications. For example, critical use cases such as drone video feeds can be prioritized over background data. Notably, the system can classify even encrypted traffic using ML models, enabling differentiated quality of service at scale. Furthermore, ML-driven dynamic network slicing supports various 5G use cases such as URLLC, eMBB, and mMTC with flexible and optimized resource allocation.

4. High Flexibility and Scalability

Built on a cloud-native, microservices-based architecture, the A5G UPF offers horizontal scalability to meet evolving traffic demands. The platform supports multi-cloud and multi-tenant deployments, and the UPF software can run on the CPU, DPU, or a hybrid setup based on deployment needs. Additionally, the solution supports multi-vendor NICs, including SmartNICs, through a platform-agnostic design that ensures broad compatibility and future-proofing.

5. Optimized Network Resource Utilization

By offloading critical UPF traffic and enabling edge computing, the solution significantly reduces backhaul network load—up to 67% less backhaul traffic. The combination of the NVIDIA Grace CPU and NVIDIA BlueField® DPU also frees up valuable CPU resources, allowing operators to run additional edge applications and services, accelerating innovation at the network edge.

Reference Architecture Highlight

The QCT S74G-2U server, powered by the NVIDIA Grace CPU and NVIDIA BlueField® DPU, supports a range of advanced 5G deployment scenarios, tailored to meet the diverse demands of modern mobile networks. These scenarios, validated through extensive testing, demonstrate how operators can optimize performance, reduce latency, and enable intelligent network functions through strategic resource allocation between CPU and DPU.

1. A5G Core software deployment on NVIDIA Grace CPU

In this scenario, the entire A5G control and user plane stack, including ML modules for traffic classification, runs directly on the high-performance NVIDIA Grace CPU. The compute power and memory bandwidth of the NVIDIA Grace architecture ensure low-latency execution and real-time decision-making. This deployment is ideal for processing requirements where very high throughput is required and UPF has more compute available for data path processing.

2. UPF Data Path Deployment on BlueField DPU

To reduce the processing burden on the CPU and reduce energy consumption for edge use cases, the UPF's data path component can be offloaded to the NVIDIA BlueField® DPU. The DPU's hardware acceleration capabilities efficiently handle high-volume data packet processing, significantly increasing throughput while preserving CPU resources for other tasks. This scenario is well-suited for data-heavy applications, such as video streaming or AR/VR services.

3. UE Flow Hardware Offload

To further optimize performance, flow offloading can be enabled for specific user equipment (UE) subnets. After the initial software-based traffic classification is completed on the Arm cores of the DPU, subsequent data packets for these flows are fully offloaded and processed in the NVIDIA BlueField® DPU hardware pipeline. This offload bypasses the software stack entirely, resulting in ultra-low latency and minimal CPU/DPU utilization, ideal for latency-sensitive applications.

4. ML-Driven Traffic Management Integration

AI/ML models are embedded into the UPF to enable intelligent traffic steering, dynamic slicing, and predictive resource allocation. The QuantaGrid S74G-2U server allows for seamless integration of ML inference engines, enabling real-time classification and steering decisions based on traffic behavior. This intelligent traffic management capability is crucial for networks aiming to support diverse services—from IoT to mission-critical applications—with fine-grained quality-of-service guarantees.

Conclusion

The integration of the NVIDIA Grace CPU and NVIDIA BlueField® DPU empowers the QuantaGrid S74G-2U server with exceptional performance, low power consumption, and intelligent traffic management, making it ideal for next-generation 5G infrastructure. Coupled with the flexibility and scalability of the cloud-native A5G UPF and QCT's robust hardware platform, this solution enables operators to significantly enhance network performance while reducing capital and operational costs. With improved efficiency and superior user experiences, QCT QuantaGrid S74G-2U server stands out as a strategic and future-ready choice for deploying high-performance, scalable 5G network solutions.

About QCT

Quanta Cloud Technology (QCT) designs, manufactures, integrates, and services cutting-edge offerings for 5G Telco/Edge, AI/HPC, Cloud, and Enterprise infrastructure via its global network. Product lines include hyper-converged and software-defined data center solutions as well as servers, storage, and network switches from 1U to entire racks with a diverse ecosystem of hardware components and software partners to fit a variety of business verticals and workload parameters.

About A5G Networks

A5G Networks is a leader and innovator in autonomous mobile core software. The company is headquartered in Nashua, New Hampshire, USA, with R&D center in India is pioneering secure and scalable 4G/5G software and packet core solutions to enable the distributed network of networks. A5G Networks' software is cloud-native and AI-native, compliant with 3GPP Release 17, and offers a 5G/4G converged core. Further, A5G Networks' AI agents enable network-wide automation, which in turn enables modern AI-driven networks and significantly reduces operational costs. A5G's software works across all the private and public cloud infrastructures as well as on all the processor architectures.

Learn More

QCT Website: https://www.qct.io

A5G Network Website: https://www.a5gnet.com

About QCT

Quanta Cloud Technology (QCT) designs, manufactures, integrates, and services cutting-edge offerings for 5G Telco/Edge, AI/HPC, Cloud, and Enterprise infrastructure via its global network. Product lines include hyper-converged and software-defined data center solutions as well as servers, storage, and network switches from 1U to entire racks with a diverse ecosystem of hardware components and software partners to fit a variety of business verticals and workload parameters.

https://go.qct.io/contact-qct-inquiry-form/

Contact Us

https://www.QCT.io

Learn More