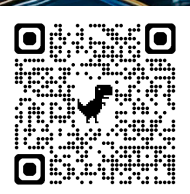


Telco Cloud Manifesto

Building an intelligent telco cloud
infrastructure for service innovation in
the mobile AI era



In partnership with



Contents

Introduction	3
Cloud infrastructure is fundamental to the success of 5G-Advanced	4
5G-Advanced promises service innovation	4
The developing role of cloud and AI in 5G-Advanced	5
Operators' spending on container platforms and AI is ramping up	7
Overcoming implementation challenges	9
Making cloud platforms carrier grade.....	9
Managing the complexity of a disaggregated and distributed architecture	9
Increasing use of AI/ML in networks demands appropriate compute resources	11
Balancing simplicity and innovation	12
Balancing the benefits of cloud-native and appliance-like experience	12
Future-ready platforms enable resource sharing and dynamic management for improved resource utilization	13
Security is a top consideration for systems leveraging agentic AI.....	13
Inherent tension between reliability and energy efficiency	14
How telco cloud delivers business-critical benefits.....	15
Full-stack deployments accelerate time to market and increase reliability	15
Improved resource optimization and cost efficiency	15
Enhanced energy savings in telco cloud	17
Case study	17
China Mobile Henan implements Telco Intelligent Converged Cloud to improve power consumption and operational costs	17
Appendix	20
Methodology	20
Acknowledgments	20



Introduction

The telecommunications industry stands at a pivotal juncture as operators transition from basic 5G connectivity to more sophisticated 5G-Advanced (5G-A) implementations foundational to the artificial intelligence (AI) era. The architecture of 5G-A is designed to both deliver AI applications and embed intelligence into common communication services. This evolution demands a fundamental rethinking of network infrastructure to support innovative service offerings with guaranteed performance.

Telco cloud architecture is the foundation for this transformation, enabling operators to deliver differentiated services with speed and latency guarantees. Leading operators are leveraging cloud-native software and AI to create compelling new services. These implementations require modifications to standard Kubernetes to support advanced networking capabilities and strict data transfer requirements in terms of input/output (I/O) performance, essential for telecommunications workloads. For many operators, full-stack deployments are popular because they balance cloud-native capabilities with operational simplicity, thereby dramatically reducing implementation timelines while improving reliability through prevalidation. These full-stack solutions run on heterogeneous computing platforms that integrate specialized processors, such as data processing units (DPUs) and graphics processing units (GPUs), alongside general-purpose CPUs. This enables efficient workload processing and reduces the size of the infrastructure footprint required.

Omdia commissioned research, sponsored by Huawei

Energy efficiency has also emerged as a critical consideration in telecom networks, requiring holistic strategies that span application, platform, and infrastructure layers. This improves sustainability through workload-aware power management, intelligent traffic steering, and energy-aware scheduling.

As 5G-A takes off, operators face the challenge of managing increasingly complex disaggregated architectures while integrating AI for network optimization. Operators that successfully navigate this transition will benefit from improved performance, enhanced mobility, and greater spectral efficiency.

Cloud infrastructure is fundamental to the success of 5G-Advanced

5G-Advanced promises service innovation

The initial 5G rollouts were based on 5G New Radio (NR). In a nonstandalone mode, 5G NR is used alongside the existing 4G packet core. This allowed telecom operators to rapidly introduce 5G services with the new air interface delivering significantly higher data speeds, reduced latency, and access to wider spectrum ranges.

However, these rollouts did not unlock the complete 5G feature set that supports ultra-reliable low-latency communications (URLLC), network slicing, and massive machine-type communications (mMTC). As a result, the first wave of 5G deployments largely failed to deliver transformative new services for communications providers. Instead, many early 5G tariff plans offered unlimited mobile data packages with higher data speeds and bundled access to popular over-the-top streaming applications. Service differentiation was primarily speed based, although speed guarantees were typically on a best-efforts basis.

The industry is now transitioning from basic 5G connectivity to more advanced implementations that can finally unlock the technology's full potential. Using a service-based architecture where network functions communicate through APIs, 5G standalone (SA) (and successive technologies) represents a fundamental redesign of the core network. The control and user plane components of the core are separated. Core network functions are cloud native, which means they can be more easily upgraded and scaled. As mobile network architecture evolves, there is an increasing emphasis on integrating intelligence, analytics, and AI to improve operational performance and enable new services.

With migration to 5G-A, telecom operators can launch more sophisticated services. Unlike the early unlimited-data 5G plans with basic speed tiers, the new 5G-A service packages offer guarantees for uplink and downlink speeds and for latency. Operators such as China

Omdia commissioned research, sponsored by Huawei
Mobile and China Unicom are also offering enhanced AI applications and industry-specific solutions with 5G-A.

Table 1 summarizes some examples from across the world of how 5G-A is allowing telecom operators globally to introduce differentiated services. Additionally, 5G-A networks are designed to serve as the connectivity layer for AI applications (in verticals including manufacturing, automotive, and utilities) by providing ultra-reliable low latency and large uplink and downlink bandwidth.

Table 1: Service differentiation with 5G-A

Service differentiation	Example
Guaranteed speed or latency	<ul style="list-style-type: none"> China Mobile Beijing offers this on themed segmented offers and, for stadium goers, utilizing dynamic guaranteed bit rate (dynamic GBR). Finland's Elisa offers a 5G fixed wireless access (FWA) network slice for remote workers and gamers. Zain KSA has launched a quality of service (QoS) prioritization plan over 5G FWA for gamers. T-Mobile launched T-Priority for first responders in the US.
Data prioritization	<ul style="list-style-type: none"> China Mobile Shanghai offers tariffs that utilize 5G-A. T-Mobile is introducing T-Mobile Secure Access Service Edge (SASE), giving customers an incremental layer of security and control when accessing certain business applications.
AI integration	<ul style="list-style-type: none"> China Mobile introduced 5G New Calling, an AI-enhanced voice and video-calling service, after the commercial launch of 5G-A. LG Uplus bundles ixi-O, an on-device AI call agent, with its 5G-A service plans. It alerts users against voice phishing calls, provides real-time translation, and answers calls on behalf of customers. The app leverages the company's ixi-GEN small language model (SLM).

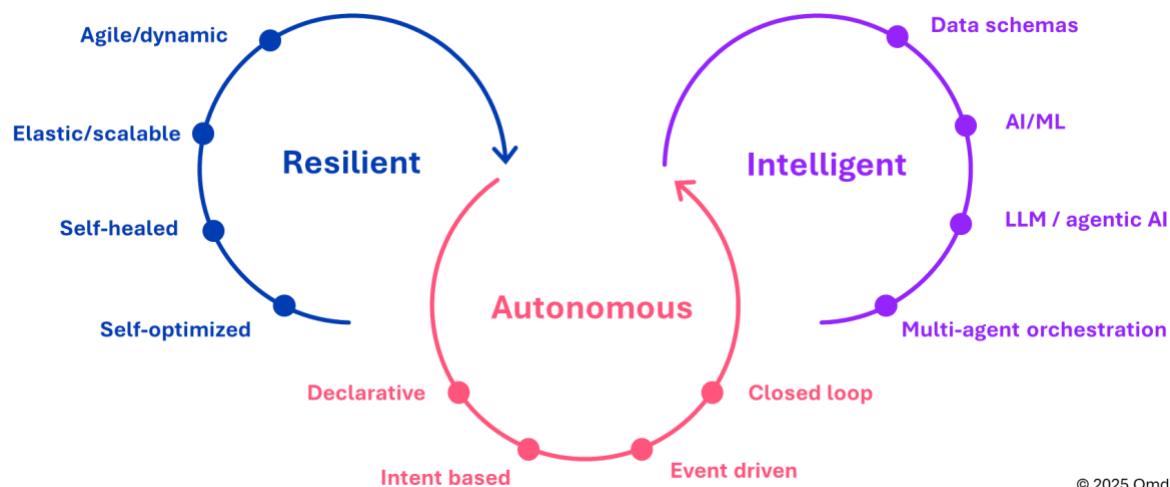
Source: Omdia

The developing role of cloud and AI in 5G-Advanced

Delivering these advanced services requires a fundamental rethinking of the network infrastructure and its operations. As shown in **Figure 1**, intelligence, automation, and resilience are the key building blocks for the new network architecture.

Omdia commissioned research, sponsored by Huawei

Figure 1: Telco cloud infrastructure requirements



Source: Omdia

Traditional networks used purpose-built hardware and monolithic systems. They lacked the flexibility, scalability, and intelligence needed to support the dynamic service environment of advanced 5G networks. Cloud technologies, specifically cloud-native design principles, microservices architecture, and containerization, enable a more flexible network infrastructure.

Cloud-native 5G networks can allocate resources dynamically, provision new services rapidly, and self-heal to ensure resiliency. These capabilities are essential for implementing features such as network slicing, which requires dynamic provisioning, guaranteed network resources, and slice-specific policies.

Offering services with performance guarantees requires closed-loop monitoring and assurance of the service KPIs and related network resources. This requires sophisticated resource coordination across the core, radio access network, and transport domains. Performance indicators are accessed as continuous streams of telemetry; legacy architectures that rely on periodic polling and batch processing are unsuitable. Cloud-native monitoring solutions such as Prometheus are designed to support more complex monitoring metrics, time series data and data models, and pull-based data collection scenarios. These open source tools are flexible, scalable, and easy to integrate with other cloud-based solutions for analytics and visualization tools.

AI is increasingly being leveraged to interpret complex data, automate actions, and enable autonomous operations. Many Tier 1 telcos, such as China Mobile, Telefónica, Telekom Malaysia, Singtel, and Vodafone, are demonstrating Level 4 autonomy across various use

Omdia commissioned research, sponsored by Huawei cases including fault management, traffic optimization, service provisioning, and energy efficiency leveraging AI and machine learning (ML).

At the same time, operators are making use of AI technologies such as large language models (LLMs) and natural language processing to enable new services such as real-time call translation, call transcription, and phishing/fraud detection. Rather than relying on public cloud providers to support these AI workloads, many operators are building their own inferencing infrastructure to meet latency and data sovereignty needs.

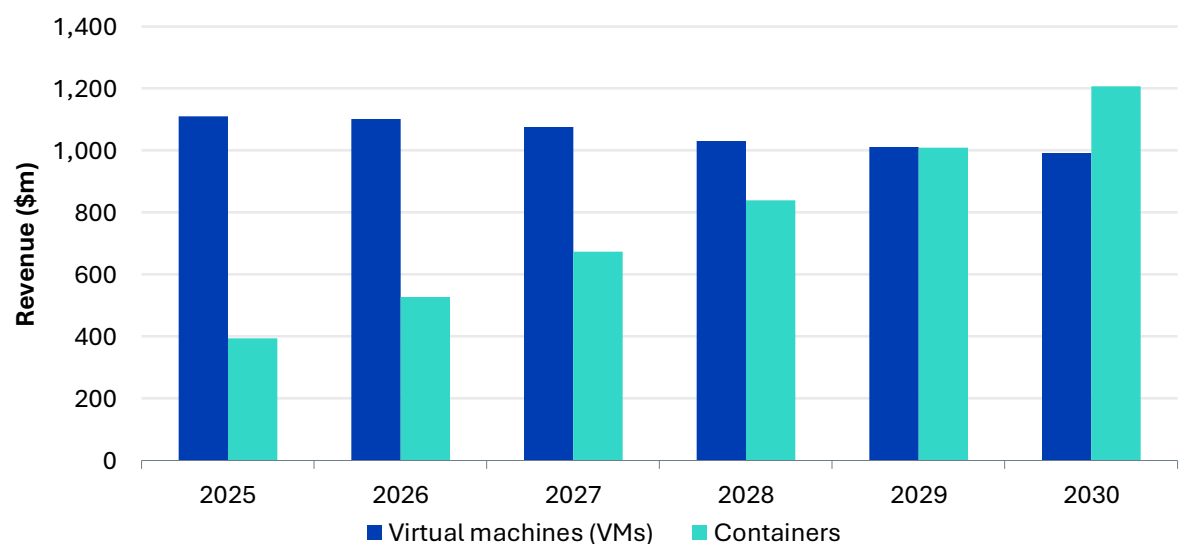
Operators' spending on container platforms and AI is ramping up

The pace of adoption of 5G core (5GC) has been disappointingly slow. According to a Heavy Reading survey of telecom operators, only about a third of respondents have 5G SA generally available. However, nearly half of the respondents indicated they will have 5G SA live by the end of 2026.

As telcos roll out 5G SA, many prefer deploying a common/combined core that can serve both 4G and 5G. With this architecture, telcos can cap any spending on existing VM-based platforms and focus their investments on platforms that support both virtualized and containerized network functions (VNFs and CNFs). Such platforms have a strong integration between the VM (OpenStack) and container (Kubernetes) environments. They offer a unified management plane and tool for deploying VMs and containers.

As **Figure 2** shows, Omdia expects spending on container-based cloud infrastructure management for network functions to exceed spending on VM-based systems by 2030.

Figure 2: Global telco network cloud infrastructure management spending, VMs vs. containers



© 2025 Omdia

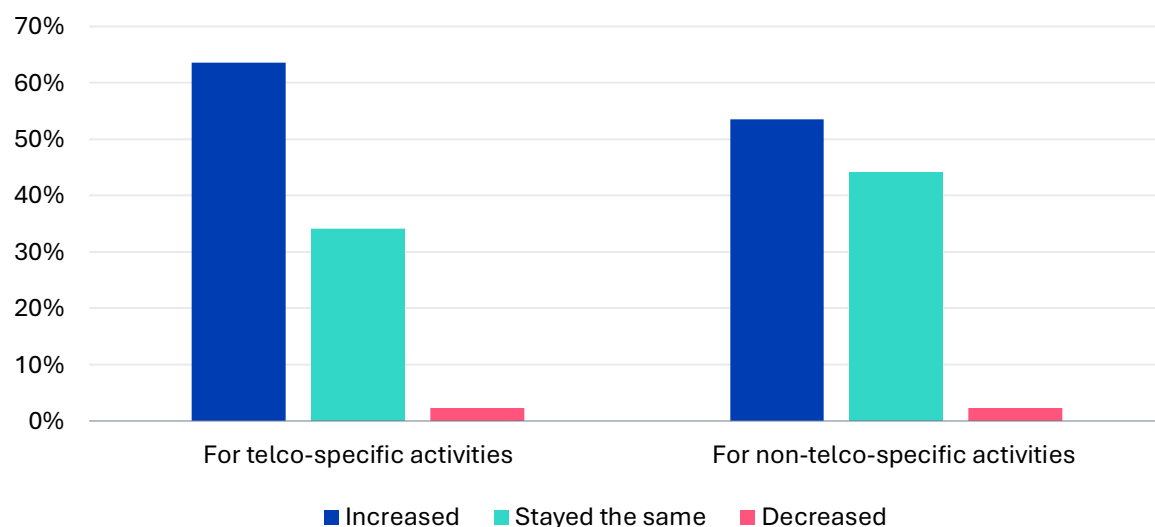
Source: Omdia

Omdia commissioned research, sponsored by Huawei

Telco spending on AI tools and capabilities is also ramping up quickly. In an Omdia survey on AI in telecom operations (see **Figure 3**), communications service providers (CSPs) indicated they have increased their use of AI for telco-specific activities, such as network management and customer care, over the past year. Though customer engagement was identified as the primary driving force for AI investments, the same survey highlighted that network management is likely to see most progress in AI enablement over the next two years.

Figure 3: Spending on AI to support telco-specific activities is growing

How would you characterize your use of AI on the job over the past 12 months?



Note: n=44

© 2025 Omdia

Source: Omdia

Within the 5G core, there are new opportunities for AI. Traditionally, data from core network functions has been used by external systems for monitoring, managing, and optimizing network performance. In 5GC a dedicated application, the network data and analytics function (NWDAF), is responsible for data-driven optimization.

Omdia's survey on core network strategies found that most CSPs are exploring ways to leverage AI for their core network functions. They expect network function vendors to embed AI capabilities within the network functions, but they are also looking to leverage NWDAF as well as external AI platforms.

The increased use of AI in telco networks is starting to influence telco cloud design choices. According to Omdia's *Telco Cloud Adoption and Vendor Perception Survey – 2025*, more than 60% of CSPs believe that their telco cloud infrastructure should also be able to

Omdia commissioned research, sponsored by Huawei host AI training and inferencing workloads. That being the case, telco cloud infrastructure must evolve beyond general-purpose computing.

Overcoming implementation challenges

Making cloud platforms carrier grade

Though CSPs can leverage many innovations from the cloud industry, such as Kubernetes (K8s), the technology must be adapted to telco environments. Telecom networks should deliver services with five-nines reliability (99.999%). High-throughput demand and latency sensitivity make a CSP's cloud infrastructure different from that of the general-purpose cloud.

One example of how telco cloud differs from generic cloud compute is the use of Multus, a container networking interface (CNI) plug-in used to attach multiple network interfaces to K8s pods. Multus can boost network performance by improving user and control plane traffic separation. Additionally, Multus is used for specialized hardware, storage-intensive applications, and multi-tenant networking.

In addition, network functions demand strict I/O performance from the underlying cloud infrastructure that generic cloud infrastructure cannot deliver. This requires modifications to “vanilla” K8s. Container plug-ins that support applications with strict latency and performance requirements include single-root I/O virtualization (SR-IOV), node feature discovery (NFD), CPU pinning, huge pages, and nonuniform memory access (NUMA). These plug-ins enable higher throughput (network I/O). Some implementations of telco cloud also support a real-time kernel to reduce latency in distributed networks.

Managing the complexity of a disaggregated and distributed architecture

The advent of 5G-A promotes disaggregated and distributed architecture that allows telcos to build networks with best-of-breed solutions. These multi-vendor implementations promise choice, but they also present significant challenges.

The decoupling of network functions from the underlying cloud infrastructure requires validation of CNFs on a cloud infrastructure. Aligning and coordinating the lifecycle management activities of the cloud infrastructure, consisting of containers and VMs, with that of the CNFs is a challenge in a disaggregated network architecture.

Telecom operators that have taken this approach spend considerable time on validating and certifying network functions from various vendors on third-party cloud infrastructure.

Omdia commissioned research, sponsored by Huawei

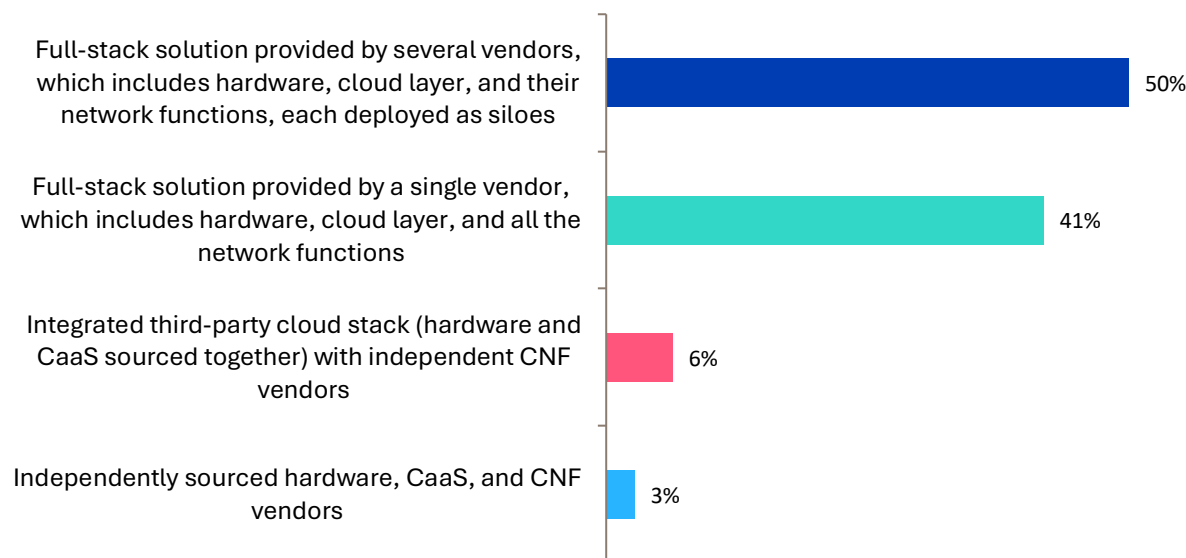
This demands extensive work, capabilities, and resources. In many cases it takes over a year to validate network functions in production environments.

Since CNFs are validated on specific infrastructure blueprints, an update to the CNFs may require a change to the underlying cloud infrastructure. Similar integrations and validations are required between the cloud software (container-as-a-service platform) and the underlying hardware.

On the customer support side, cross-layer fault finding and demarcation of responsibilities between the different vendors becomes challenging. The idea of “single handshake” or having a single entity responsible for delivering the desired service-level agreements still resonates with many telcos. The offer of a full end-to-end support contract covering hardware, cloud infrastructure, and applications is attractive for CSPs with thinly stretched technical staff. In Omdia’s *Service Provider Core Networks Survey – 2025*, when asked which cloud technology strategy CSPs would use, 50% of respondents said they would deploy full-stack solutions from several vendors in silos.

Figure 4: Fifty percent of CSPs would use a full-stack solution from several vendors

What type of cloud technology stack will your company use for 5G core?



Note: n=111

© 2025 Omdia

Source: Omdia

Omdia commissioned research, sponsored by Huawei

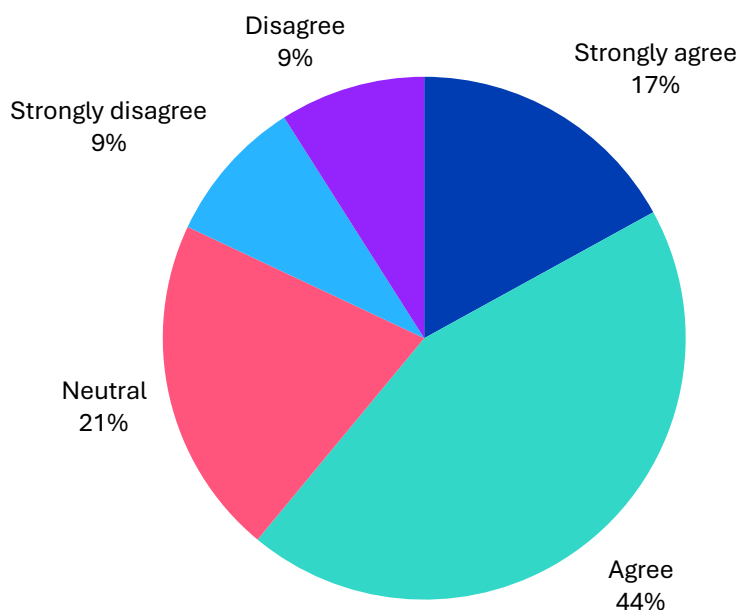
Increasing use of AI/ML in networks demands appropriate compute resources

Mobile networks leverage both traditional and generative AI in many ways. CSPs are exploring AI opportunities across service creation, orchestration, and assurance. AI can help by generating configuration templates for specific service requirements (such as latency, bandwidth, and reliability) based on the intent declared by the user. AI can determine the optimal resource distribution to support a service instance and automatically configure resources to meet service characteristics.

For service assurance, AI can continuously monitor service performance and decide on specific actions such as scaling or descaling resources in near-real time. Such closed-loop systems can be implemented with the help of generative AI and agentic AI. Several AI agents could be used to observe the performance of a service, identify the root cause in the case of service degradation, suggest appropriate actions to remedy the issues, evaluate the actions, and actuate the desired changes.

LLMs are foundational to agentic AI. Training and inferencing such LLMs is computationally intense. As shown in **Figure 5**, a recent Omdia survey of telecom operators found that the majority (60%) believe that support for ML training and inferencing workloads is becoming a critical factor in their telco network cloud infrastructure decisions.

Figure 5: Support for ML training and inferencing workloads is becoming a critical factor in telco network cloud infrastructure decisions



© 2025 Omdia

Source: Omdia

Omdia commissioned research, sponsored by Huawei

Balancing simplicity and innovation

Balancing the benefits of cloud-native and appliance-like experience

There are many architectural models available to CSPs for rolling out their cloud-native infrastructure. Some Tier 1 operators are choosing horizontal implementation, where network functions from different vendors are deployed on a common cloud infrastructure. Though this model gives telcos more control, flexibility, and choice over their infrastructure, it also demands significant investments in building tools, skills, and operational capabilities. For telcos with limited resources, managing cloud-native infrastructure in tandem with older technologies can become an operational headache leading to long lead times and high integration costs.

A low-risk approach involves deploying a prevalidated full-stack solution for containerized network functions. This model provides better and more integrated technical support from network function vendors. It delivers the benefits of cloud with an out-of-the-box or “appliance-like” experience.

A full-stack solution bundles centralized and automated lifecycle management of CNFs and cloud infrastructure. It provides a simplified upgrade path with a single vendor responsible for maintaining cloud software, hardware, and network functions. It is critical for telcos choosing a cloud deployment model to understand the cost of upgrades. This will be linked to upgrade frequency, time, and complexity.

A unified automation framework is necessary to manage frequent updates to CNFs, CaaS infrastructure, and so on. The ability to jump Kubernetes versions and migrate directly to the desired version reduces the number of upgrade cycles required.

For telco cloud, which requires many custom enhancements, close collaboration between the hardware and software layers is also important. Optimizing collaborations between different components ensures that tasks such as traffic offloading from CPU to acceleration cards and smartNICs are performed efficiently.

Vendors stacks are fine-tuned for optimizations across the OS layer, container orchestration platforms, and the network applications. This improves the overall resource utilization when a prevalidated stack is being deployed. However, such deployments come at the cost of limited scalability and choice.

Omdia commissioned research, sponsored by Huawei

Future-ready platforms enable resource sharing and dynamic management for improved resource utilization

For most telecom operators, the implementation of virtualization technology for 4G core infrastructure was not very different from their appliance-based rollouts. VNF deployments typically accompanied dedicated hardware resources, which were statically allocated, and offered no dynamic scaling.

Many initial 5GC deployments seemed to replicate this approach. Operators chose to set up separate or isolated environments for containerized network function (CNFs), which were generally overprovisioned for peak capacity. This architecture is simpler to implement, but it results in significantly underutilized compute, storage, and memory resources.

Sharing cloud resources between VMs and containers involves complex integrations because they demand different networking plug-ins, storage-provisioning methods, and security controls. Resource sharing also requires coordination of scheduling, scaling, and resource allocation across heterogeneous or separate runtimes for containers and VM.

As telcos evolve their cloud infrastructure, they must take a holistic view of their existing deployments, founded mainly in VM-based infrastructure (such as OpenStack), the need for more dynamic resource allocations for containerized workloads, and additional compute demanded by AI. When the AI inferencing hardware is being introduced, creating an integrated infrastructure where GPU (primarily used for inferencing) and CPU (used for other general computing) resources are managed within a single unified cloud environment can offer several benefits. Such heterogeneous cloud infrastructure should enable better hardware (compute, storage) management to support network functions and AI. It also reduces the overheads for managing separate resource pools.

Unified telco cloud platforms deliver a common management plane for integrated monitoring, control, and visibility of VMs and containers. Enhanced schedulers provide common frameworks for scheduling AI, network functions, and IT-type workloads. Additionally, the resource schedulers should be capable of multidimensional resource management supporting CPUs, GPUs, and other resources.

Security is a top consideration for systems leveraging agentic AI

Enabling more sophisticated 5G SA services requires telcos to modernize their IT systems. In many instances, operators are complementing existing conversational AI implementations in OSS/BSS with agentic AI capabilities. In some cases, AI agents might

Omdia commissioned research, sponsored by Huawei generate code to be used in other systems. Such autonomous code generation can introduce vulnerabilities through

- Lack of security review before execution
- Potential injection of malicious instructions through prompts
- Unintended logic flaws that create exploitable vulnerabilities

Therefore, autogenerated code must be tested in a sandbox or digital twin before being used in the production network.

Inherent tension between reliability and energy efficiency

Alongside resource efficiency, telco cloud implementations must also address energy efficiency. Though AI has amplified the need for improving data center power consumption, telcos have a bigger challenge at hand.

A telco data center consists of many idle server nodes that are not allocated for actively running workloads. A key reason is overprovisioning of compute or other resources because CSP networks are designed to meet future capacity growth. High availability in architecture is a must for cloud-native deployments; network functions are designed for N+1 and 2N redundancy. Server nodes supporting core network functions offer guaranteed capacity and must provide additional headroom for increasing traffic load up to a certain threshold before the performance indicators are affected. Overall, a 5GC cluster, for example, has idle server nodes in addition to unallocated CPU cores. At the same time, utilization of allocated CPU cores varies between busy and nonbusy hours.

For telcos that are trying to address energy consumption in core network infrastructure, a key starting point is to improve the reporting of relevant energy KPIs. There are a range of point solutions available today to help telcos improve their sustainability goals and reduce power consumption of cloud infrastructure. CSPs can choose between smarter orchestration and workload placement, adjusting performance (CPU throttling, dynamic scaling) for off-peak power saving, use of hardware offload/accelerators, and consolidating network functions on fewer resources. Point solutions target cloud infrastructure, but telcos need to take a holistic approach—across application, platform, and infrastructure—when addressing energy efficiency in 5G-A networks.

Omdia commissioned research, sponsored by Huawei

How telco cloud delivers business-critical benefits

When they are properly designed with carrier-grade infrastructure, resource optimization, and robust security, telco cloud implementations offer compelling benefits that have a direct impact on operational efficiency and service agility.

Full-stack deployments accelerate time to market and increase reliability

Full-stack 5GC deployments enable telecom operators to balance cloud-native capabilities with operational simplicity. The full-stack approach integrates network functions, cloud middleware, operating systems, and hardware into a prevalidated, cohesive package managed through centralized automation.

Comprehensive integration delivers five critical advantages for operators:

- Deployment cycles shrink dramatically because prevalidated stacks eliminate the complex integration testing typically required between components. This reduces implementation timelines from many months to weeks.
- Upgrade management becomes more streamlined through unified automation frameworks that coordinate updates across all stack layers simultaneously. This minimizes the service disruptions that plague fragmented approaches.
- Version compatibility challenges are effectively eliminated through the ability to migrate directly to desired Kubernetes versions during update cycles, bypassing the incremental upgrades required in horizontally deployed environments.
- Reliability improves substantially through comprehensive pretesting across the full stack, ensuring all components work harmoniously and reducing production incidents related to component interactions.
- Performance optimization emerges naturally because vendor stacks are fine-tuned across all layers from operating systems through container orchestration platforms to applications. This delivers performance advantages that are not available in generic deployments.

Improved resource optimization and cost efficiency

As discussed in the previous section, the diversity of applications supported in a telco's network environment has increased, leading to the need for careful design of the cloud infrastructure. Some workloads are hosted as VMs; others have evolved to run over containers. A practical approach is to share the deployed resources, but doing so is a

Omdia commissioned research, sponsored by Huawei complex process. Cloud platforms that bring this convergence as an inbuilt capability offer several benefits. They provide a smooth evolution path for legacy applications, thus protecting a telco's investment in its cloud infrastructure.

In addition, diversity is also seen in the performance requirements of various workloads. Some telco applications are closer to generic IT; therefore, they do not place special processing demands on cloud infrastructure. Such applications are managed well with CPUs, the general-purpose processors.

Then there are core network workloads that impose slightly higher performance requirements. The control plane functions are less sensitive to latency; their performance requirements are easily met through CPUs. User plane functions introduce complexities in the overall computing architecture. They need to integrate specialized processors to support traffic offloading and acceleration. This is important for making sure CPU resources are not overburdened and to meet the demanding latency and throughput requirements.

A telco cloud requires sophisticated hardware architecture that efficiently supports diverse workloads through heterogeneous computing. Using traditional computing architectures for AI-driven use cases can be inefficient. Such approaches, where communication between GPUs is forwarded through CPUs, can create unnecessary bottlenecks, adding to the overall inference latency. A unified approach integrates specialized processors (network processors, DPUs, and GPUs) alongside general-purpose CPUs to ensure optimal workload placement and processing efficiency. It does so by leveraging architectures that introduce a high-speed interconnect bus to allow xPU communication directly without waiting for CPU processing.

This unified approach offers several advantages:

- Optimized workload processing by matching tasks to appropriate silicon
- Accelerated packet processing through SmartNIC and DPU interface cards that offload network functions while freeing CPU resources
- Enhanced AI capabilities through dedicated GPU resources for high-throughput inference essential for network optimization
- Efficient resource utilization through optimized component collaboration
- Reduced infrastructure footprint through consolidation of diverse workloads on unified hardware

Omdia commissioned research, sponsored by Huawei

Beyond the unified management and scheduling of general-purpose and AI computing hardware, another key consideration is an AI enablement platform. This platform provides capabilities for AI model management, scheduling, and acceleration—supporting agile AI application development and rapid deployment—further strengthening the telco cloud’s support for AI-driven service innovation.

Enhanced energy savings in telco cloud

Telco cloud implementations require holistic energy-efficiency strategies spanning application, platform, and infrastructure layers. This comprehensive approach delivers significant sustainability improvements through

- Workload-aware power management that dynamically allocates resources and consolidates workloads
- Intelligent traffic steering that routes data through energy-efficient processing paths
- Hardware-level optimization using purpose-built acceleration cards
- Integrated thermal management that responds to actual utilization patterns
- Energy-aware scheduling algorithms that balance performance with power consumption.

By addressing energy efficiency systematically rather than through isolated component optimizations, operators can simultaneously meet sustainability targets and reduce operational costs while maintaining service quality, a critical advantage in an increasingly energy-conscious telecommunications landscape.

Case study

China Mobile Henan implements Telco Intelligent Converged Cloud to improve power consumption and operational costs

China Mobile Henan is part of China Mobile, a major telecommunications services provider in China. The company provides mobile communications in Henan province using 5G and 5G-Advanced. With a focus on delivering a differentiated service experience for its customers, it became one of the first operators in the country to commercially test 5G-A in 2024. In early 2025, China Mobile Henan worked with Huawei to rollout the 5G New Calling service, which allows real-time interactions such as in call voice translation and transcription. It leverages AI technologies to deliver these differentiated service

Omdia commissioned research, sponsored by Huawei experiences to its customers, embedding intelligence into both its cloud-native networks and service delivery.

Challenges

Incorporating AI technologies can have a significant impact on the telco's compute resources. In China Mobile Henan's case, it leveraged LLMs and digital twins for automated network management (fault, configuration, accounting, performance, and security). As it did so, it needed more general-purpose compute and specialized computing hardware, GPUs, and other AI accelerators to train as well as inference models. Choosing an appropriate telco cloud solution was important to balance the compute costs and energy consumption. GPUs and other AI accelerators are expensive, and utilization and energy efficiency is often suboptimal, adding both financial and sustainability pressure.

Solution

Huawei's Telco Intelligent Converged Cloud (TICC) helped China Mobile Henan in achieving its goal of implementing AI-driven solutions across multiple layers of operations and realizing notable reductions in power consumption and operational costs while maintaining high service quality. The two collaborated to address the following:

- Enhancing CPU utilization through intelligent scheduling and energy optimization
- Introducing an AI-enabled platform to accelerate intelligent agent applications and AI model deployment
- Actively engaging in industry standards and ecosystem building to drive unified frameworks

Outcome

As a result of the collaboration, China Mobile Henan has achieved the following benefits:

- TICC provided intelligent frequency scaling, server down-clocking, and sleep modes. With AI-driven workload sensing, CPU frequency is dynamically adjusted, cutting power consumption per server by up to 20%.
- Through intelligent scheduling, defragmentation, and resource aggregation, TICC reduced server fragmentation by 20%. TICC also incorporates an intent-based engine that automates server power on/off, achieving both energy savings and deployment efficiency.
- With panoramic visibility and digital-twin monitoring, TICC tracks power usage across tens of thousands of devices in real time. Additionally, fault localization time is shortened from hours to minutes, reducing operations and management costs by 30%.

Omdia commissioned research, sponsored by Huawei

- In the pilot results, China Mobile Henan also noticed an overall reduction in the power consumed by each server (approximately 40W). This is achieved with a combination of techniques available through TICC. According to China Mobile Henan, the overall solution is expected to save 1.5GWh annually, cutting 805 tons of carbon emissions, equivalent to planting 50,000 trees.

Omdia commissioned research, sponsored by Huawei

Appendix

Methodology

This white paper is based on interviews with experts from Henan Mobile and China Mobile, Huawei, and OpenInfra. It leverages Omdia's existing research and knowledge on multiple topics such as telco cloud, telco software and AI, and core networks as input.

Acknowledgments

We would like to thank our contributors:

Jiexuan Du, Henan Mobile, Department of Network Administration

Longbing Tan, Henan Mobile, Department of Network Administration

Beichen Xu, Henan Mobile, Department of Network Administration

Zhiyu Fan, China Mobile, Department of Network

Ildiko Vancsa, OpenInfra, Director of Community

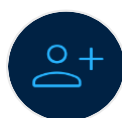
Bing Xu, Huawei Cloud Core Networks, Senior Engineer

Peng Ding, Huawei Cloud Core Network, Senior Expert

Xiaobin Wang, Huawei Cloud Core Network, Director of Marketing Execution Dept.

Huaiqi Wang, Huawei Cloud Core Network, Director of Product Portfolio & Lifecycle Mgmt. Dept.

Xuezhe Guo, Huawei Cloud Core Network, Marketing Specialist



Inderpreet Kaur, Senior Analyst, Telco Cloud & Network Automation
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 TechTarget, 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.

Get in touch

www.omdia.com
askananalyst@omdia.com



Copyright notice and disclaimer

The Omdia research, data, and information referenced herein (the "Omdia Materials") are the copyrighted property of TechTarget, Inc. and its subsidiaries or affiliates (together "Informa TechTarget") or its third-party data providers and represent data, research, opinions, or viewpoints published by Informa TechTarget 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 TechTarget 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 TechTarget 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 TechTarget 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.