The edge cloud: an agile foundation to support advanced new services

White paper

As telecoms networks continue to evolve towards 5G and the Internet of Things (IoT), there is a growing requirement to provide low latency services, as well as to increase capacity. While higher and more flexible network capacity is being achieved with cloud technologies, the industry has so far adopted a centralized architecture, inspired by the IT cloud.

While this meets the needs of telecom workloads that have migrated to the cloud and has provided extra capacity at a lower overall cost, it is less suited to the needs of new applications with larger data payloads and low latency.

Moving compute capacity closer to the traffic, at the edge of the network, helps to address this shortfall. Called the edge cloud, this approach creates a layered network architecture that combines centralized and edge data centers. The edge cloud has the flexibility to take communications into the 5G era and support advanced, low latency applications that will transform our way of life.

Contents

Executive Summary

Cloud computing is transforming the telecoms landscape, offering Communications Service Providers (CSPs) flexibility and lower overall costs, plus high service availability to meet the growing demands of a fully connected world.

Current telco cloud implementations follow traditional IT cloud architecture, with traffic transported to a small number of centralized data centers that handle the telco workload and applications.

To complement this approach, compute power should also be implemented closer to the origination of traffic, at the edge of the network. Known as edge cloud, this approach offers many advantages by better supporting high bandwidth and/or low latency applications.

Edge cloud is built on a layered network architecture, combining centralized and edge data centers, with capacity distributed according to the requirements of the workloads. In addition, common management and orchestration layers make management easier for CSPs.

Placing compute capabilities close to the antennas aggregates the processing from base stations to achieve greater efficiencies. Edge cloud complements the centralized and regional data center infrastructure approach embodied in the Cloud Radio Access Network (RAN).

Beyond this, edge cloud also enables new low latency services and makes it possible to process large amounts of data at the edge of the network without needing to transport all that data to centralized data centers.

Cloud native edge infrastructure will be essential to enable the successful implementation of 5G and to support new, advanced vertical use cases powered by network slicing capabilities.

The Nokia vision for the 5G era is described by its Future X network architecture. This reference architecture for a distributed network is the foundation for the edge cloud.

The changing network landscape

While today, nearly everyone is connected, we are moving quickly into a future in which nearly everything will also be connected. 5G offers the promise of ultra-broadband connectivity everywhere, meeting the extreme performance demands of new applications by creating a seamless fabric of interconnected intelligence. The Internet of Things (IoT) will extend traditional business models and drive network investment and improvements to operational processes.

The opportunities created by 5G are immense. They will have major implications for the way we live in and organize our societies, manage our industries, control transport and healthcare and many other aspects of the modern world.

These market drivers will require increasingly diverse services and applications with a wide range of demands for latency, privacy, security, bandwidth, cost savings and reliability.

Many CSPs have gone some way to meeting these demands by adopting cloud technologies, predominantly for their core networks. Cloud technologies enable a high degree of automation that reduces manual work, provides elasticity that enables networks to grow and shrink according to demand, and delivers better user experiences.

Implementing cloud computing has so far mostly addressed functions such as the IP Multimedia Subsystem (IMS), Packet Core and Virtualized Customer Premises Equipment (vCPE), evolving from a simple virtualization approach into a more advanced cloud-native design. Implementing a cloud-native software architecture allows unpredictable demand to be met more efficiently. The network can adapt instantly to traffic fluctuations while resources are matched to the needs of different services. To achieve all this flexibility and efficiency requires effective management and orchestration solutions that implement a high level of automation.

Network Functions Virtualization (NFV) and Software Defined Networking (SDN) are vital building blocks for 5G core networks, enabling the use of pooled infrastructure for higher utilization and simplicity. Without cloud capabilities, CSPs will not be able to serve growing traffic demand efficiently or run advanced capabilities such as 5G network slicing.

The edge cloud advances these principles still further. As demand for more responsive services grows, for Machine Type Communications (MTC), and with the deployment of 5G, low latency becomes a clear requirement. Furthermore, as data traffic grows, transport networks will come under increasing pressure as large traffic volumes are transported to core locations.

What are edge clouds?

The edge cloud places compute capabilities close to where traffic originates, at the edge of the network. This decreases latency and removes the need for all traffic to run over the full transport network, optimizing the use of network resources. Edge cloud resources will also enable new applications, such as virtual reality, augmented reality and autonomous driving, that need to be run close to the data sources.

A variety of network functions will run at the edge. With the transformation to a cloud-native design, core network functions like Evolved Packet Core (EPC) and 5G Next Generation Core (NGC) will run the control and user planes at the edge.

However, it is not feasible to simply move all workloads to the edge of the network. Instead, a balance of centralized and distributed compute resources in a layered architecture across the network between the core and edge will enable workloads to be placed where they can best support the required service and traffic profiles.

How edge clouds are built

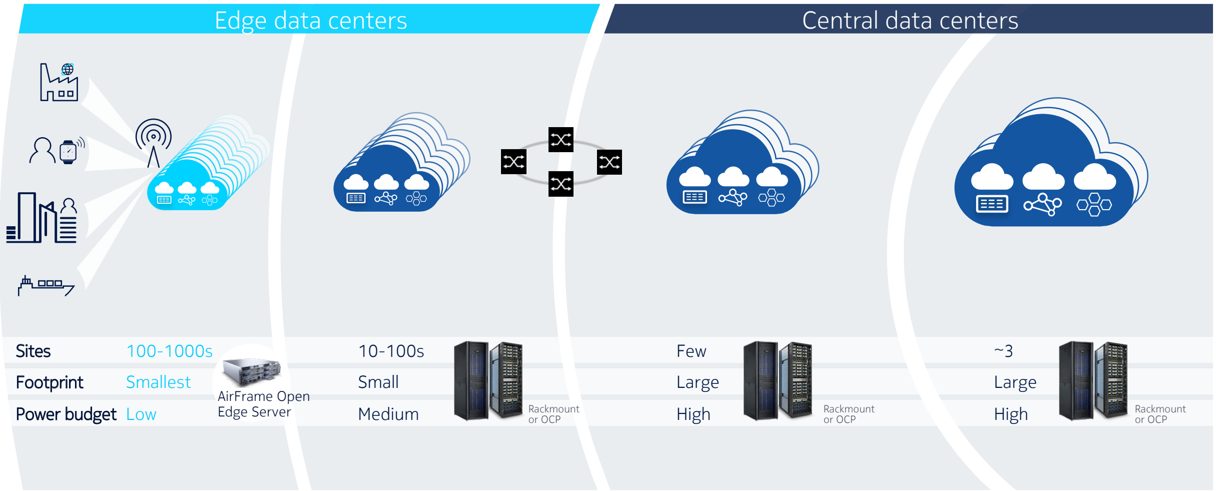
A scalable data center fabric provides connectivity within the edge cloud and towards external networks, in particular the cloud-native core network.

The building of an edge cloud is governed by a variety of requirements including:

- The need for physical space and power for the compute hardware
- The need to support real-time workloads
- The need for management tools for distributed data centers and clouds
- The need for orchestration capabilities for distributing and managing workloads in hundreds of small clouds.

A typical layered cloud architecture, encompassing centralized and edge data centers is shown in figure 1. In the radio network the edge data centers fit well on base station sites, which can act as aggregation sites for a number of base stations.

Figure 1. Layered architecture in which applications and VNFs are distributed across layers based on their requirements and characteristics



Sites	100-1000s	10-100s	Few	~3
Footprint	Smallest	Small	Large	Large
Power budget	Low	Medium	High	High

Cloud RAN: a single, orchestrated cloud for radio access

Perhaps one of the most prominent examples of a network function running at the edge is the Cloud RAN. Telco cloud architecture in the radio access domain needs to provide real-time services as well as meet the specific requirements of RAN functions. Cloud RAN requires powerful data center infrastructure with real-time optimization and very low latency at the cell sites.

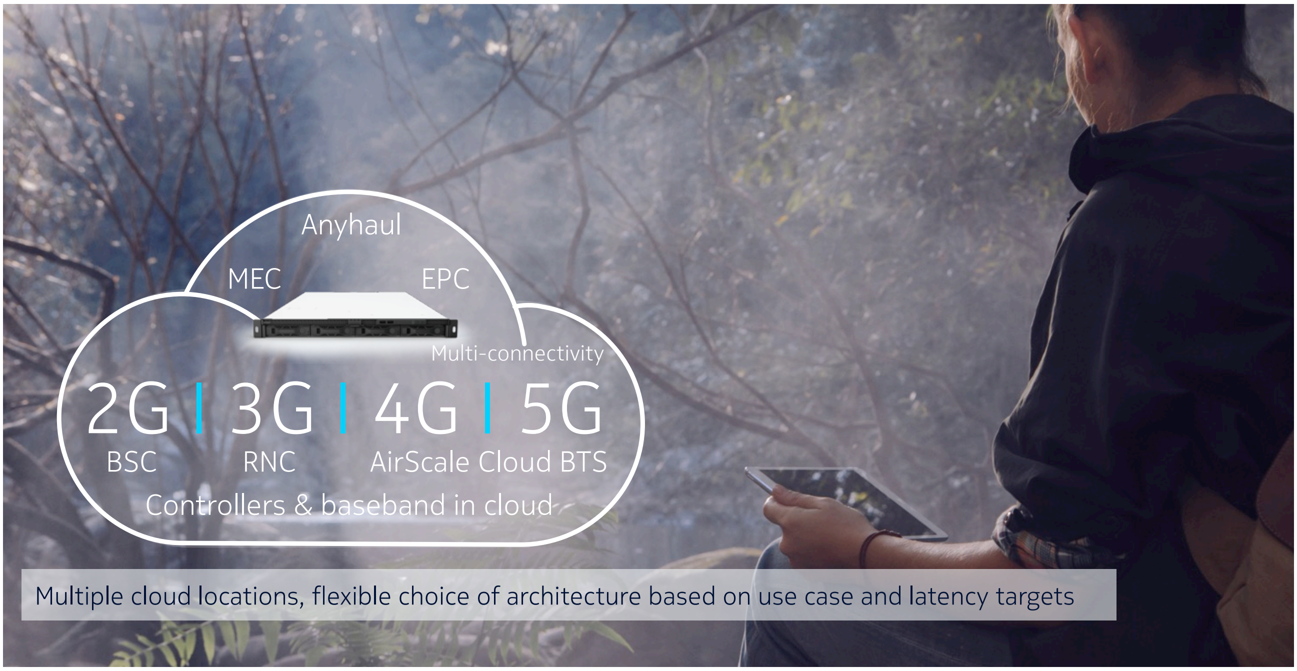
Nokia Cloud RAN architecture supports radio functions in large, centralized data centers, as well as in smaller, distributed data centers, for example, in aggregation points or cell sites, and at the edge, capable of delivering real-time services to end users. In the Cloud RAN, all the layered cloud components become an effective single, orchestrated cloud that enables CSPs to flexibly deploy virtualized RAN functions.

The influences on where to deploy functionality can include performance requirements (such as bandwidth or latency), the CSP's available assets (such as suitable aggregation sites and transport infrastructure) and CSP deployment strategy and business case.

The fully scalable solution further allows optimal integration for cloud application servers such as Multi-access Edge Computing (MEC) which can run on the same servers as Cloud RAN. This opens up APIs to new business opportunities, applications, services and plug-ins, seamlessly integrated into the RAN and using information from real radio conditions.

No matter what the architecture choice, multi-connectivity capability on the Cloud RAN platform is needed to enable multiple radio technologies to collaborate as one system. In practice, multi-connectivity Packet Data Convergence Protocol (PDCP) splits the data packets over multiple radio technologies (LTE, Wi-Fi, unlicensed, 5G) to enable capacity load balancing across layers. This improves agility, time to market, effective radio resource utilization and helps CSPs to dynamically exploit and coordinate all their radio assets according to radio conditions or applications used.

Figure 2. Cloud RAN based on layered compute infrastructure meets the extreme performance demands of 5G and IoT


Anyhaul
MEC
EPC
Multi-connectivity
2G | 3G | 4G | 5G
BSC
RNC
AirScale Cloud BTS
Controllers & baseband in cloud
Multiple cloud locations, flexible choice of architecture based on use case and latency targets

MEC: an edge cloud applications enablement platform

Technologies like 4G and 5G will take advantage of cloud capabilities, including edge cloud, to power simple and complex use cases with consistent reliability and performance. The use cases range from providing immersive video experiences to tens of thousands of people in stadiums, to powering the fourth industrial revolution with connected robots, to enabling smart cities by connecting and analyzing millions of IoT devices.

To meet these needs, Multi-access Edge Computing, or ETSI MEC, is currently the only specified framework to enable applications in a distributed telecommunications network. MEC is currently being brought into the 5G architecture by 3GPP. The 3GPP focus is on specifying the necessary enablers that allow application functions to use network services.

MEC acts as an application enabling framework and provides platform services to orchestrate applications based on technical and business parameters. It defines APIs and services to bring the application and network worlds together. The target is to make it easy to develop applications by providing common practices and APIs.

MEC Application Enablement enables authorized applications to access the CSP network, lets them run within the network and consume network and context information via the APIs. It can run on an add-on network element, sharing the virtualization resources with other network functions or applications, or on a standalone host with virtualization resources dedicated to MEC.

MEC Application Enablement provides:

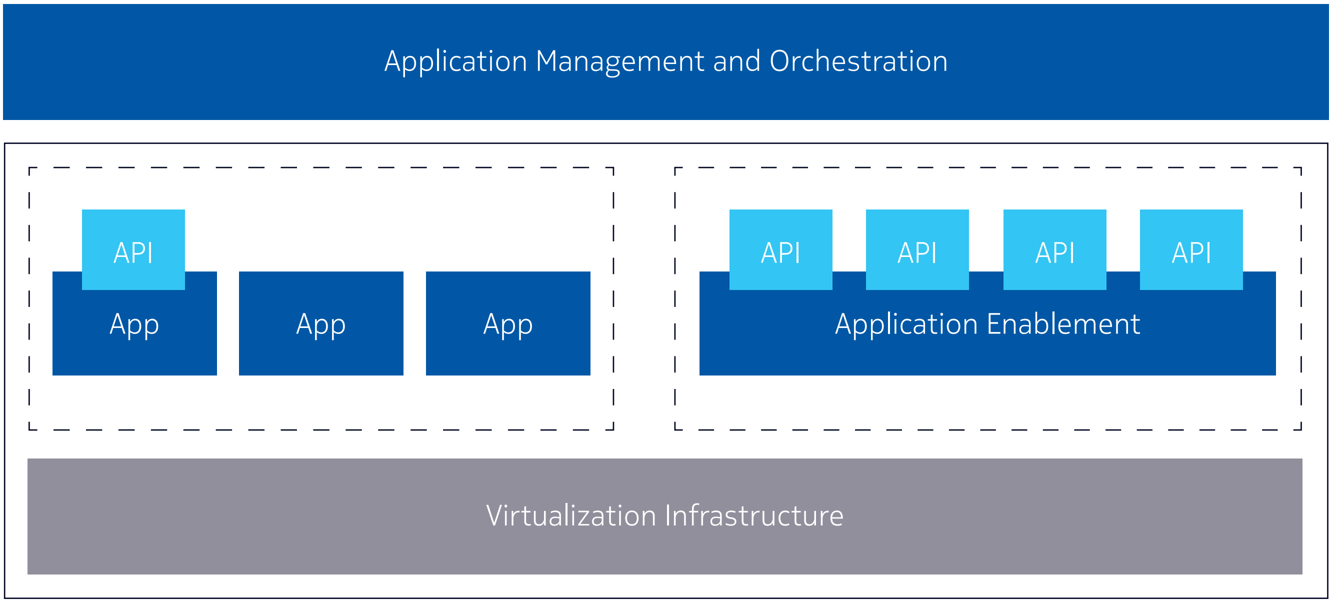
- Registration, announcement, discovery and notification of services
- Authentication and authorization of applications that provide and use services
- Communication support for services (query/response and notifications).

With MEC, services can be both offered and consumed at the most appropriate locations within the network. For example, a sports stadium could use MEC to provide high quality video streams from the event to spectators. An airport could use MEC for advertising, location and augmented reality. An industrial plant could use MEC for video surveillance and as an IoT gateway for connecting IoT devices. A campus or conference center could offer local services to residents and visitors. Distributed MEC is deployed close to the actual venue or within the venue, for example, an enterprise or stadium.

The customer experience is greatly improved by MEC's ability to deliver real-time mobile services that use context information and location awareness to create a high degree of personalization. These services can also be more responsive because of the ultra-low latency achieved by locating computing resources near to the point of use. Popular and locally-relevant content can be delivered from exactly where users consume it.

The reduced latency, high bandwidth and local context can also support the needs of critical communications and IoT applications that demand robust and highly responsive connectivity.

Figure 3. ETSI MEC acts as an application enablement platform and is being brought in by the 3GPP as part of its next generation network architecture


Application Management and Orchestration
API
App
App
App
API
API
API
API
Application Enablement
Virtualization Infrastructure

Designing open edge computing hardware and software

The edge cloud has different requirements for data center hardware than traditional centralized data centers. At the edge, base station site real estate space is often severely restricted and conventional data center hardware will not fit into same-sized cabinets used for base station equipment. Furthermore, electrical power supplies at base station sites tend to be low capacity. This all means that edge data center hardware must be more compact and more energy efficient than conventional data center hardware.

These issues can be resolved by applying modern data center design, such as Open Compute Project (OCP), which is claimed to be the world's most efficient data center hardware.

Hardware acceleration is essential in edge data centers to enable them to meet the low latency requirements of many applications. A prime example is 4G/5G baseband processing, which is very processor intensive. Artificial intelligence, especially machine learning, is another clear example of an application that will benefit from hardware acceleration.

Additionally, cloud infrastructure software must be real-time, highly available and scalable to support the performance requirements of the edge.

Nokia AirFrame open edge server is the first x86 portfolio built and tailored to fully support edge cloud deployments. It combines an ultra-small footprint with a real-time OpenStack distribution built to provide the performance and low latency required by Cloud RAN and network slicing.

Nokia AirFrame open edge hardware is complemented with cloud infrastructure software that supports scalability and performance in the 5G era. Nokia Airframe Cloud Infrastructure for Real-time applications is an NFV solution for managing both virtual machines and containers in an environment comprising a huge number of distributed clouds. Being an Open Platform for NFV (OPNFV) compatible OpenStack distribution, this open solution is enhanced to run with real-time performance, hardware acceleration, telco-grade operability and minimal cloud overhead even in very small data centers.

Conclusion

The edge cloud will be fundamental to the evolution of networks to enable CSPs to support new services and applications, opening several new vertical business opportunities. By starting now to prepare and implement edge infrastructure, CSPs can begin to address these opportunities and be ready for the next evolutionary steps of their networks, including 5G.

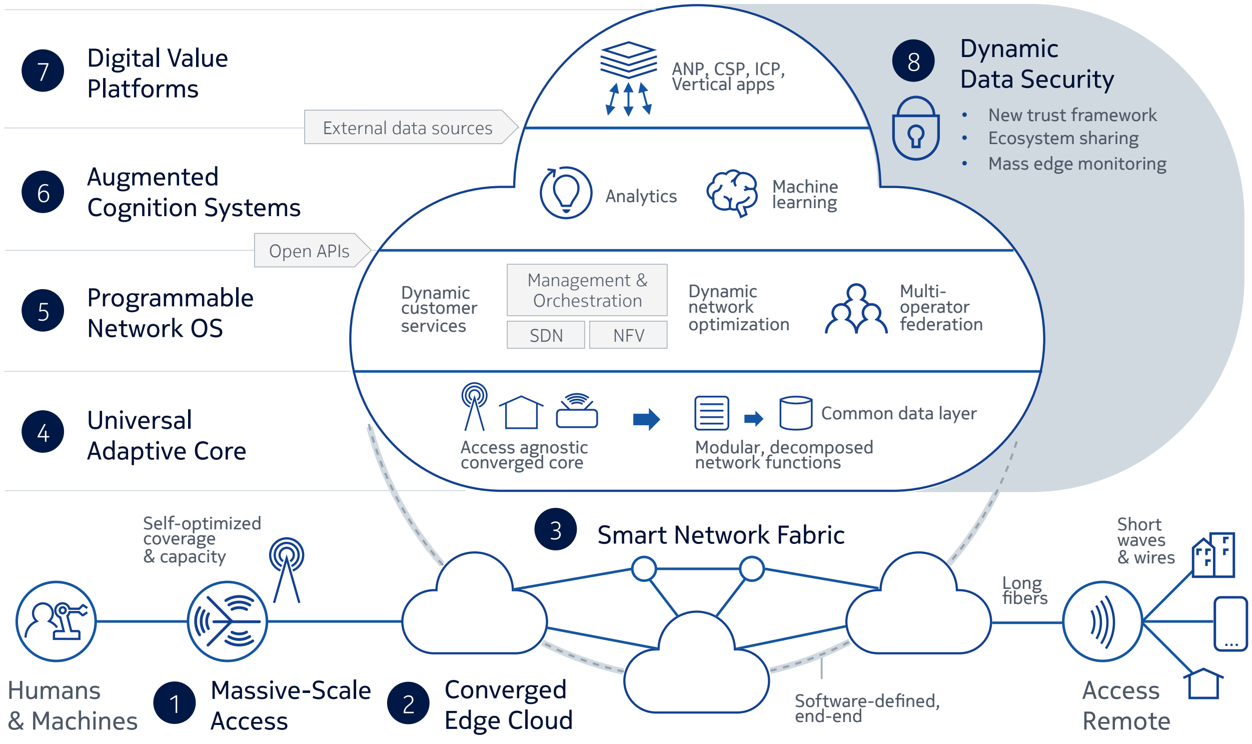
The most prominent example of edge cloud deployment is the Cloud RAN, which supports radio functions in large, centralized data centers, as well as in smaller, distributed data centers. Cloud RAN acts as a single orchestrated edge cloud to provide real-time services as well as meet the specific requirements of RAN functions.

Currently being brought into the 3GPP next generation network architecture, ETSI MEC provides a platform that enables applications to use edge cloud capabilities. With MEC, services can be both offered and consumed at the most appropriate locations within the network.

Nokia AirFrame open edge cloud infrastructure directly addresses the requirements for building compute capacity at the network edge and will be a fundamental asset to build the networks of the future.

Nokia Future X network architecture is a comprehensive vision for the 5G era. Future X is Nokia's reference architecture for a distributed network. It is the foundation that defines how Nokia executes edge cloud.

Figure 4. The Future X Network architecture showing its evolution to the converged, cognitive, cloud-optimized network



7 Digital Value Platforms
External data sources
6 Augmented Cognition Systems
Open APIs
5 Programmable Network OS
4 Universal Adaptive Core
ANP, CSP, ICP, Vertical apps
Analytics
Machine learning
Dynamic customer services
Management & Orchestration
SDN
NFV
Dynamic network optimization
Multi-operator federation
Access agnostic converged core
Modular, decomposed network functions
Common data layer
8 Dynamic Data Security
New trust framework
Ecosystem sharing
Mass edge monitoring
Self-optimized coverage & capacity
3 Smart Network Fabric
Short waves & wires
Long fibers
Humans & Machines
1 Massive-Scale Access
2 Converged Edge Cloud
Software-defined, end-end
Access Remote

Abbreviations

API	Application Programming Interface
CSP	Communications Service Provider
EPC	Evolved Packet Core
IMS	IP Multimedia Subsystem
IoT	Internet of Things
MEC	Multi-access Edge Computing
MTC	Machine Type Communications
NFV	Network Functions Virtualization
NGC	5G Next Generation Core
OPC	Open Compute Project
OPNV	Open Platform for NFV
RAN	Radio Access Network
RNC	Radio Network Controller
SDN	Software Defined Networking
vCPE	Virtualized Customer Premises Equipment

About Nokia

We create the technology to connect the world. Powered by the research and innovation of Nokia Bell Labs, we serve communications service providers, governments, large enterprises and consumers, with the industry's most complete, end-to-end portfolio of products, services and licensing.

From the enabling infrastructure for 5G and the Internet of Things, to emerging applications in digital health, we are shaping the future of technology to transform the human experience. networks.nokia.com

Nokia Oyj
Karaportti 3
FI-02610 Espoo, Finland
Tel. +358 (0) 10 44 88 000

Document code: SR1712020305EN (June)