FutureWorks

5G - a System of Systems

for a programmable multi-service architecture

Contents

Executive Summary

The future of mobile communications is likely to be very different to today's experiences. New uses will arise, many not yet conceived, creating novel requirements that communications networks must meet flexibly and cost effectively to support operator profitability and the wider ecosystem. How we travel, how we experience our environment, how we control remote environments, how the infrastructure supports us and how we produce goods will all be changed by 5G connectivity.

A very diverse set of Machine Type Communication (MTC) services is emerging with significantly different characteristics from today's dominant human-to-human (H2H) traffic. Traffic volume is increasing and service requirements are becoming more diverse, posing severe challenges for current mobile network architecture from a technical and business point of view. These need to be addressed and tackled to pave the way for the Programmable World in the 5G era.

A frequently asked question is: 'can the current Evolved Packet Core (EPC)/ System Architecture Evolution (SAE) efficiently support the services of the 5G era?' The answer is clearly no. Existing LTE and EPC architecture is not able to support low latency services, for example, especially when they also need full mobility. In addition, the explosion of mobile broadband and massive deployment of MTC - aka Internet of Things (IoT) - will drive network complexity and related costs. This can endanger a sustainable business.

Whether steering robots in a factory, delivering UHD video, or connecting healthcare sensors, 5G networks must support far more uses than 4G networks that were primarily designed for delivering high speed mobile broadband.

While it may be straightforward to build a separate system for each of these requirements and use cases, the real challenge is to develop 5G as a single system of systems that can meet all these requirements invisibly from the user's perspective. Therefore, 5G is much more than a new radio system and will require an overhaul in the network to keep complexity and cost manageable. The total cost of ownership is reduced by using resources much more efficiently. Full flexibility and programmability are achieved by the underlying principles of cloud, virtualization - referred to as network function virtualization (NFV) - and software defined networking (SDN).

To support new business models the traditional 'one size fits all' network architecture needs to change to a 'flexible per service' paradigm. 5G networks will enable functions to be offered as a service, that is as a Network-as–a-Service. If all network elements from access, core, OSS to security and analytics are virtualized and sliced out as one integrated 'service', it should be possible for an operator to create an instance of an entire network virtually, relying on whatever underlying infrastructure is available for the defined geography. Using the power of programmability, the operator can customize such a 'network instance' for any industry enterprise, whether automotive, healthcare, logistics, retail or utilities.

The new 5G System of Systems will be built on a strong foundation with an adaptive and programmable multi-service architecture based on the Cognitive and cloud Optimized Network Evolution, CONE [2].

The road to this 5G network architecture transformation is not trivial and market needs are not homogeneous around the globe. Nokia believes that the new 5G architecture should be introduced in phases.

1. The first phase solves imminent business needs to boost mobile broadband capacity and leverage existing LTE deployments by tightly connecting 5G radio access to the 4G network via dual connectivity.
2. The second phase introduces an optimal architecture for supporting diverse use cases like massive MTC, critical MTC and extreme mobile broadband. Furthermore, radio agnostic core networks will be introduced to provide the long term convergence for radio and fixed access. Network slicing allows network resources to be tailored to the needs of the application and capability of the device.

The phased approach will allow operators to leverage existing deployments to provide higher data rate, more capacity in the near term and at the same time, introduce a future-proof network architecture to support new use cases and services in the longer term. Thus, it will enable operators to have the right tools to unlock new business potential by offering new services via their 5G network.

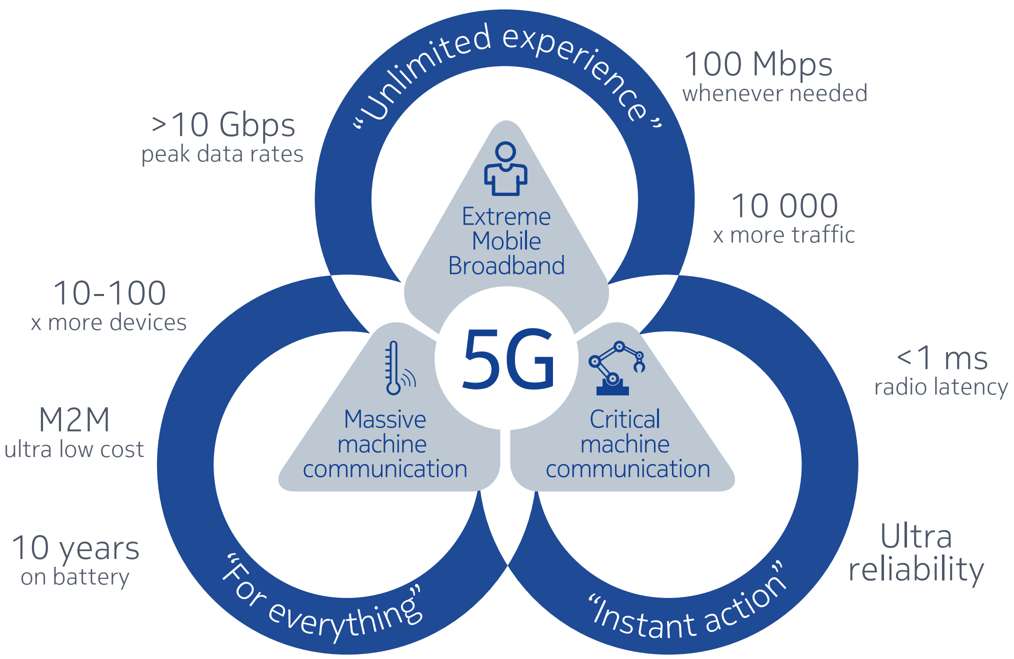

Figure 1: 5G will need to support a wide variety of different uses and business models

New business potential for operators

5G network architecture must efficiently support new use cases and operator business models. It must also have minimal network TCO and device costs [1].

TCO can be minimized by improving the network's energy efficiency, while features such as session and mobility on demand enable better use of network resources and extended device battery life.

Innovative solutions can help operators to improve their profitability. Such innovations lay in three areas that enable operators to make money from their network investments:

Network performance

The new performance level of their networks enables XXL and low latency broadband services such as HD and UHD in the home and on the move, but also virtual reality services for the business world. These "Connectivity+" business models provide new opportunities through guaranteed high service levels not only with end users, but also with content and other service providers.

Providing superior QoE in mobile broadband is extremely important since it will be the major revenue source in the 5G era. Dynamic Experience Management (DEM) enables monetization of the quality of experience for B2B and B2C models and ensures extremely efficient resource utilization.

In addition to throughput, low latency will enable new use cases such as vehicular communication (which requires many-to-many connectivity), industrial control and virtual reality. Seamless service continuity should also be offered during periods of full mobility by supporting gateway relocation. Low latency is needed for use cases like automotive, Tactile Internet and industry automation. Some of these, such as V2X and industry automation, also require ultra-reliable connectivity.

Network data

The millions of transactional and control data points produced by the network can be leveraged to enable entirely new services that benefit from contextual real-time and non-real-time data. Operators can broker this information to different industries such as providers of augmented reality services, traffic steering systems (for example provided by municipalities), factories, logistics and utilities.

Network slices

Dedicated virtual sub-networks, the network slices, can be marketed as Network as a Service (NaaS) which provide exactly the functionality needed for different industries and their diverse use cases. For example, the functionality needed for connecting huge numbers of consumer health sensors is completely different to that required for high quality UHD video delivery to TV sets. The network slicing provides the flexibility to run multiple network instances (slices) on the same physical network infrastructure. It also allows operators and vendors to enable new business models by offering NaaS. This can be extended from the core network down to radio access. The flexibility behind the slicing concept is a key enabler to both expand existing businesses and create new revenue sources.

Key functionality of the programmable multi-service architecture

To enable the new and diverse 5G use cases, the system architecture needs to provide a set of new functionality as summarized in Figure 2.

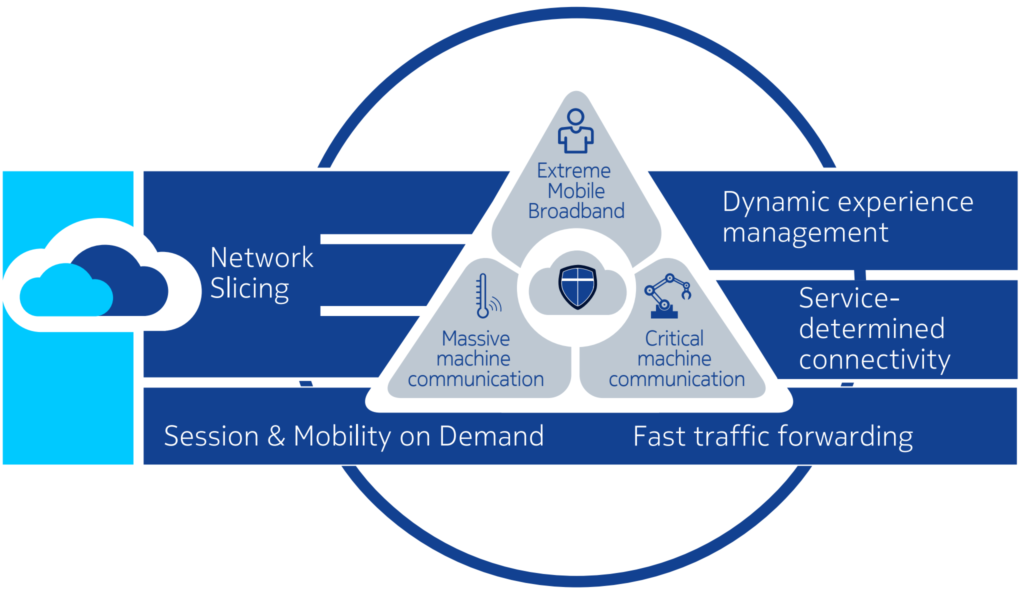

Figure 2: New functionality of the 5G architecture

Network slicing

5G will not only be a ‘new RAT family’ with its radio access network but its architecture will expand to multiple systems by providing a common core for multiple radio technologies (cellular, Wi-Fi, fixed), multiple services (mobile broadband, Massive MTC and critical MTC) and multiple network and service operators. This is enabled by NFV and SDN technologies which allow systems with a high level of abstraction. In 5G, we envision that networks will be further abstracted with the concept of network slices, where each slice is tailored for a specific use case. Figure 3 gives an overview of network slicing as enabler of new Network-as-a-Service business models.

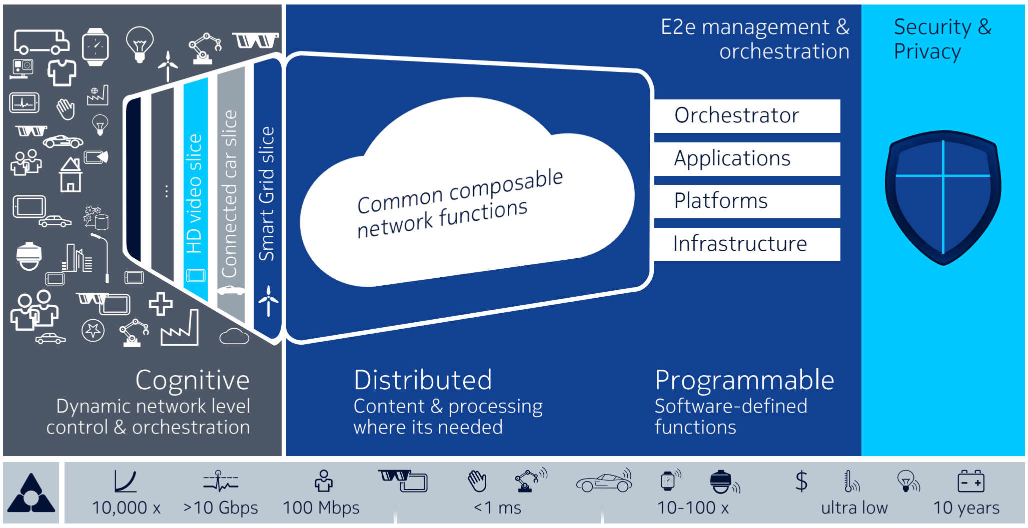

Figure 3: Network Slicing enabling Network-as-a-Service

Dynamic experience management

User experience management to enable a good QoE is very important for, but not limited to, mobile broadband services. The LTE bearer centric Quality of Service (QoS) architecture is unable to differentiate between application sessions served by the same bearer unless dedicated bearers are used for different categories of applications. As long as services can be differentiated via 5-tuples filter (Source and Destination IP address/range, Source and Destination Port number/range, and Protocol ID of the protocol above IP), EPS bearers can be used. But this is not possible for all services, either because they do not differ in the 5-tuples or the 5-tuples need to change dynamically (e.g. when embedded content in an HTML page is loaded from a server) which makes an update in the device and gateway almost impossible. So, to make LTE applications aware, the EPS QoS architecture requires a major overhaul that impacts the device and network protocol stacks from upper Network Access Stratum (NAS) layer down to the lower Medium Access Control (MAC) layer thus evolution of EPC is not suitable to tap the full potential of dynamic QoS/QoE management.

Service-determined connectivity

Support is needed for new use cases that require re-locatable low latency and high reliability (multi-connectivity) services. In LTE the continuous optimization of the IP anchor point (i.e. relocation of the IP anchor) is not supported. But this is necessary to be able to offer the shortest and best path for routing user plane traffic and enabling low latency services (e.g. critical MTC), while at the same time support seamless service continuity during mobility. In LTE, the Packet Data Network (PDN) connections with a local gateway are possible for Local IP Access and traffic offloading. But, if the user moves beyond the serving area of the local gateway, there is no support for continuity for the

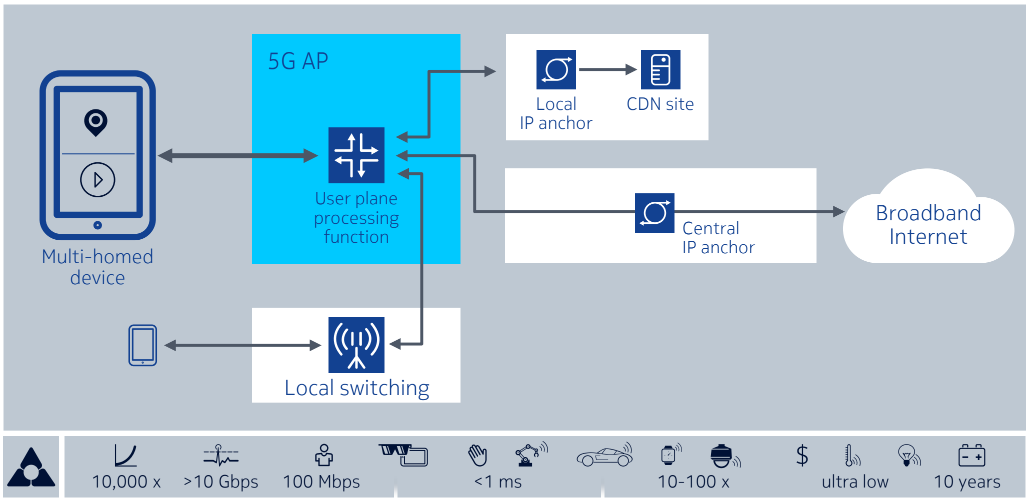

Figure 4: Service determined 5G connectivity model

corresponding service. In addition a p-t-p connection to a (central) gateway is not appropriate for some use cases like V2X that require support for low latency with full mobility and many-to-many connectivity. Figure 4 shows the different networking models that need to be enabled by 5G. Another example is to provide simultaneous access to Internet services through a central gateway and access to local CDN site through a local gateway.

Fast traffic forwarding

In order to support low latency services, the gateway and application should be close to the radio, a technology referred to as Mobile Edge Computing (MEC). Depending on the location of the communicating devices, switching happens either at the radio or in an aggregator cloud. Furthermore, service continuity must be maintained when the communicating devices are fully mobile.

Session on demand

When 5G architecture is introduced, the 'always on' principle needs to be modified into a 'session on demand' principle. LTE does not allow the flexibility to introduce "signaling only sessions" for massive numbers of IoT devices. The 'always on' user plane connectivity can result in battery drain for such devices, e.g. animal tracker devices with button-sized battery that require maximum optimization. At the same time, for smartphones, the 'always on' principle can be retained for efficient mobile broadband services. Modifying such principles would require an overhaul of mobility management and session management procedures in LTE/EPC.

Mobility on demand

Offering 'mobility on demand' makes efficient use of network resources. Depending on the application needs and device capabilities, the network determines the right level of active and idle mode mobility to be assigned for a certain device.

Flexible mobility consists of two components: one for managing mobility of active devices and a second for tracking and reaching devices that support a power-saving idle mode. The assigned mobility may range from one extreme, beginning with no 'active mode' mobility, with no support for idle mode (typical with today's Wi-Fi access) to the other extreme with full support for active and idle mode mobility as applied in 2G/3G/4G. Different levels of flexible mobility bridge the gap between these extremes, allowing for independent assignment of idle-mode mobility on a per-device basis, and active mode mobility on a per-application basis. Flexible active mobility is made possible by supporting flexible IP anchoring and enabling mobility only when needed.

As well as the new key functions outlined above, the 5G communication system will differ in some fundamental ways to current networks:

Separation of access-specific and access-independent functions

In current EPC architecture, dependency between radio and core does not allow operators to evolve their radio and core networks independently. The 5G architecture should address this in such a way that the core network is agnostic to the (radio) access, enabling radio and core to evolve independently. This will also enable a long term convergence with Wi-Fi and fixed access. The result will be an access-agnostic core where basic functions such as QoS, session management and security principles are decoupled from the underlying access technology.

Ultra-reliability

Critical services such as public safety communication, health and safety, industry automation and automotive require (extreme) high-availability and reliability. This is implemented by native HetNet, where multi-connectivity improves robustness, reliability and data throughput. Multi-connectivity or single connectivity are selected depending on the type of service to improve network efficiency. In addition, some of the above use cases require an isolated mode of operation at the radio access network when there is a backhaul failure towards the core network.

Network resilience features should be designed into the 5G network. Network elements must also fulfill high-availability requirements (99.9999 percent) and to recover from hardware failures. Pooling of elements (hot and cold standby) and load balancing between core elements improve system reliability and guarantee network service availability, even if a single node fails.

Figure 5 zooms into the 5G components, namely the 'virtualized 5G network functions' and 'Shared Data Layer (Data Repository)', of the overall end-to-end architecture vision as outlined in [2].

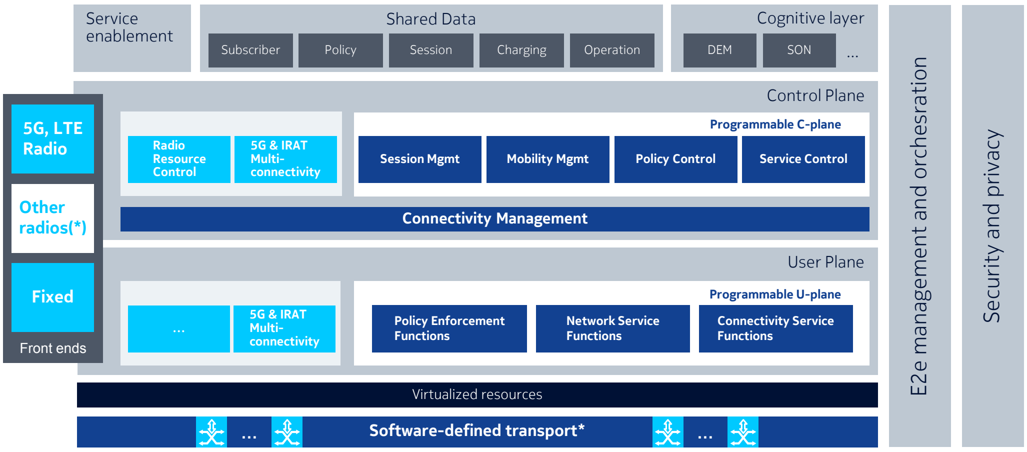

Figure 5: 5G network architecture functions

The baseline for the architecture, shown in Figure 5, is the virtualization of 5G network functions and the separation of the control and user planes, both in the access (light blue) and non access (dark blue) domains.

Mobility management, session management, service control functions and dynamic policy control will benefit from common data layer(s). The connectivity management functionality supports the flexible definition of connectivity models going beyond point-to-point services.

On the user plane, seamless integration of Network Service Functions - also referred to as service functions e.g. firewall, Deep Packet Inspection (DPI), parental control that form the service chain - on top of the basic connectivity and policy enforcement function, removes the separation between the network service functions that are traditionally placed in the SGi LAN and the connectivity service. However network service functions can be deployed independently in the SGi LAN or seamlessly integrated to the basic connectivity and policy enforcement functions if required. Network Service Functions must be applied to the packets and/or frames in the downlink and uplink path based on the service needs and this is defined as the Flexible Service Chain. Furthermore, flexible service chaining can take full advantage of edge clouds. Policy enforcement functions will be enhanced with active QoE management to ensure a premium user experience even in congested network areas.

When it comes to the radio access, what can or cannot be virtualized depends on front-haul availability. The radio architecture needs to be flexible to cope with different front-haul deployments. Typical radio functions that can be virtualized include the Radio Resource Control / Management and the multi-connectivity (intra-5G and 5G-LTE dual connectivity). A key feature of 5G, multi-connectivity can be at the higher and/or lower layers.

SGi LAN refers to the operator controlled network "behind" the mobile access. The mobile access ends at the P-GW. An overview of the 3GPP architecture can be found in 3GPP TS 23.401 [3].

Will 5G be a revolution or evolution?

To provide the required data rates, latencies, robustness and connectivity, novel technologies are required for 5G. Most of these technologies are characterized by their flexibility and ability to adapt to different scenarios and use cases. In both radio and core network telco cloud, virtualization and Software Defined Networking will enable network-wide programmability, elasticity and scalability. Network orchestration enables automation across network components through centralized management of network resources.

The 5G system will be a mix of well-proven 4G and novel technologies. Some existing technologies will continue to evolve, while new technologies will be introduced with 5G.

5G will natively enable diverse use cases with an extreme range of requirements. Nokia's view is that 5G has to be built on top of LTE and EPC and that a healthy balance of evolution and revolution will ensure investment protection for operators and vendors while innovative technologies will enable a future-proof 5G network.

There are two main technical reasons for evolution:

1. Mobile broadband (the main use case) represents the majority of revenues and data volume today and in the future.
2. We define 5G as a 'System of Systems'. This means that 5G is a multi-RAT network, which natively combines LTE and new 5G technologies (in addition to other access types like Wi-Fi or fixed).

However, while retaining the proven foundation, Nokia believes that new network architecture is needed for several key reasons.

When considering the key functionalities needed for the programmable multi-service architecture and try to implement in the current one we see that the loss of flexibility would lead to an ineffective and expensive network. The substantial increase of mobile broadband and massive deployment of MTC (IoT) traffic in the coming years needs proper scalability and programmability. The abstraction of network functions from the underlying hardware is essential. It is expected that network functions will run as software components on top of operators' telco cloud systems rather than using dedicated hardware components. In LTE, virtualization is applied in a box-driven way. In 5G, virtualization should be used as the underlying principle for the new architecture design. The breaking down and reassembly of network functions to better use NFV technologies for best scalability and agility is vital.

Separation of the control and user plane functions is the basic SDN principle, which enables dynamic allocation of user plane resources at the best place for a given service. Some example use cases that will benefit from a flexible gateway allocation depending on the service are low latency services and/or Content Delivery Network (CDN) which require the gateway close to the radio access while the best option for basic Internet access may still be a central gateway.

‘Rome wasn’t built in a day’ – the phased approach

The first commercial 5G deployments are expected in 2020 [1]. By then, most major operators will have completed the majority of their LTE roll outs. This will include the use of all licensed and unlicensed spectrum, with LTE and 5G tightly integrated. At this point the capacity demand will reach a tipping point for mobile broadband and IoT which will require a more cost effective solution. How is it possible to cope with such an ambitious timeframe?

Nokia foresees a phased approach for 5G radio and architecture development and deployment considering market requirements and different use cases as well as standardization roadmaps such as 3GPP. The ‘road to 5G radio and architecture’ is envisioned as follows:

Phase 1

5G radio will be optimized for extreme mobile broadband use, as shown in Figure 6. In Phase 1, the architecture is centered on EPC and deployment of new 5G radio access technologies. The new radio access technologies will be affected by the availability of higher frequency spectrum assigned for mobile use by the World Radio Conference (WRC).

In order to meet the 2020 target for first commercial deployments, the 5G radio specification should be ready by 2018. This is achieved by introducing 5G radio in a dual connectivity mode with LTE as an anchor. In this scenario, the 5G radio can have user plane connectivity to EPC either directly or via the LTE eNB. 5G radio access will have control plane connectivity towards EPC via the LTE eNB only as there is no need for 5G radio in the dual connectivity mode to have direct control plane connectivity to the EPC. The architecture considered for this phase is in line with NGMN network architecture option 1

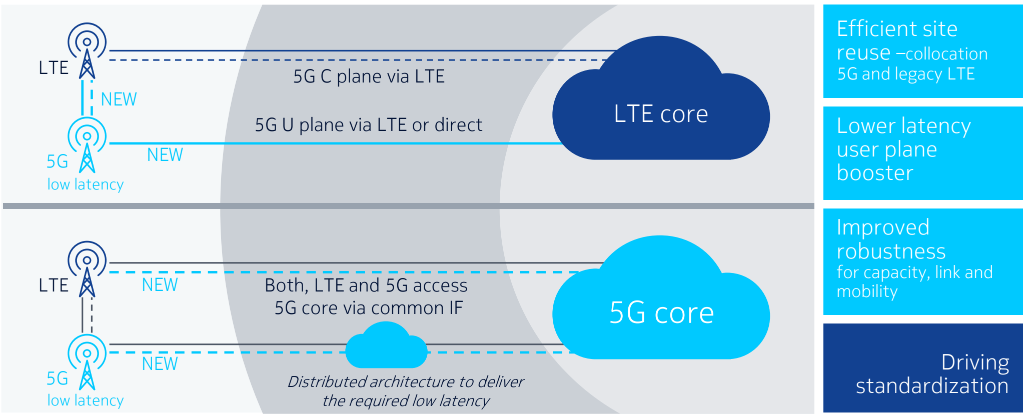

Figure 6: Phased approach for 5G radio and architecture development and deployment

from the user plane perspective. EPC is the core for both LTE and the new RAT in dual connectivity mode. The minimum feature set to ensure forward compatibility towards phase 2 should be introduced during phase 1 to support for new use cases and services in the longer term. [1]

Phase 2

The end-to-end 5G system introduces a new 5G core network and standalone 5G radio access without the need for an LTE anchor. In this phase, the end-to-end 5G system will meet the requirements of critical MTC use cases by supporting lowest latency with full mobility and highest reliability. Important mobile broadband enhancements will provide best quality of experience (QoE) to end users and also efficiently support massive MTC.

Operators will benefit from the new architecture's flexibility with features like mobility on demand and session on demand that will affect the Total Cost of Ownership (TCO). Providing the best point of attachment for static and nomadic users (and support mobility only when they really are on the move) without going through the central gateway will save transport network resources. Session on demand will save user plane resources when serving massive MTC, e.g. signaling only or connectionless data transmission.

This phase foresees a new 5G core network architecture and corresponds to NGMN network architecture option 3 [1]. The 5G network architecture will consider NFV, distributed Cloud and SDN as underlying principles; the functional decomposition and reassembly of today's network boxes will enable the best network architecture to be designed in a fully virtualized environment.

Nokia roadmap to the 5G System of Systems

Figure 7 shows the 5G architecture roadmap to 2020 to develop and deploy the key elements of the System of Systems

Nokia expects 5G radio and architecture to be standardized in 3GPP. The standardization of 5G radio and architecture has been kicked off in fall 2015. Phase 1 of the RAN specification should be ready by 2018 to pave the way for commercial deployment in 2020. ETSI ISG on Mobile Edge Computing (MEC) that started in 2014 plays a role in the 5G System of System especially for supporting low latency use cases as has the platform for edge computing.

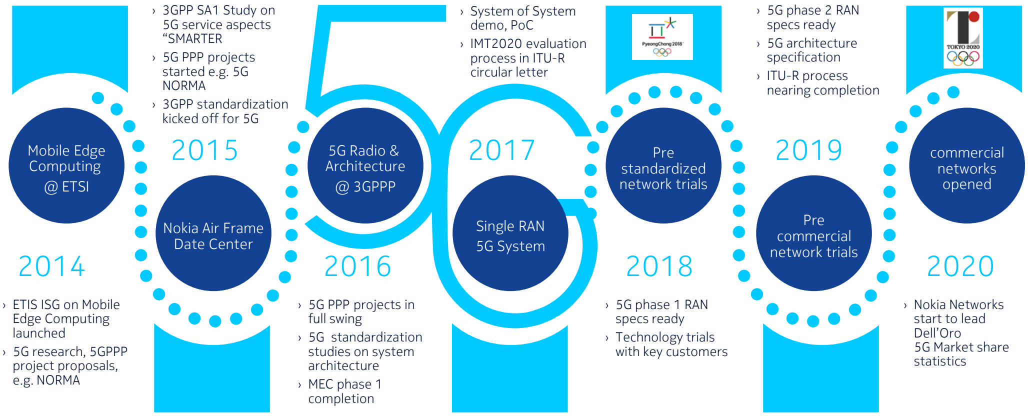

Figure 7: Architecture roadmap towards the 5G System of Systems

Nokia Networks collaborates with leading universities in Europe, the USA and China to make 5G technologies and architecture as viable and future-proof as possible. It also runs several 5G related industry projects in Europe, chairs the 5G-PPP association and is running 5G innovation projects with major operators including e.g. China Mobile, Deutsche Telekom, Korea Telecom, NTT DoCoMo, SK Telecom and Verizon.

Nokia already showcased various 5G technologies and is a driving force to make 5G happen:

- Nokia launched its new AirFrame Data Center Solution which helps operators bring their data centers into the cloud.
- During Mobile World Congress 2015 Nokia Networks showed novel 5G technologies such as mmWave and cmWave, demonstrating performance benefits through multi-connectivity between LTE and 5G.
- In the European 5G Public Private Partnership (5G-PPP) Nokia Networks takes a leading role in Europe's largest 5G projects. Nokia Networks is one of the driving forces behind the METIS-II project, providing the technical coordinator of the project and leading the overall 5G RAN design and the spectrum work in METIS-II. This engagement is complemented by the 5G NORMA project coordinator role taken by Nokia Networks. In this project a novel radio multiservice adaptive network architecture will be developed. Furthermore, Nokia Networks leads the architecture and system design related work packages in the 5G-PPP projects mmMAGIC and FANTASTIC-5G. The engagement in these projects positions Nokia Networks to ensure a holistic 5G radio and network system design.
- Research collaborations: Nokia maintains tight links with leading academic institutes in the area of 5G research, e.g. TU Kaiserslautern, TU Dresden (which was recently named by Fierce Wireless as among the top five universities leading worldwide in 5G research); Nokia is also a leading industry partner of the 5G research at NYU (also ranked among the top 5 universities for 5G).

- Nokia has various bilateral collaborations with leading operators worldwide encompassing radio system topics and architecture aspects and proofs of concept.
  - NTT DOCOMO on 5G radio technologies (mmWave and below 6GHz)
  - SKT for virtualized RAN as well as for cmWave radio access
  - KT on “Programmable World” (5G and IoT LTE-M)
  - China Mobile and Deutsche Telekom on 5G radio access technologies
  - Verizon on 5G sandbox verifications with the aim to target commercial deployments of fixed 5G access solution to connect to the home
- Nokia will work with its partner to provide pre-standardized 5G mobile trials during a major sports event in 2018 in South Korea, and then strive for true commercial network availability in 2020 during another global sports event in Japan.

Conclusion

This white paper outlines the need for a new 5G network architecture as well as the path for such a transformation.

- Whether delivering UHD video, steering robots in a factory or connecting healthcare sensors, 5G networks must support far more uses than 4G networks that were primarily designed for the single use case of delivering high speed mobile broadband.
- To support the stretched performance targets of 5G and the diverse range of services, the traditional ‘one size fits all’ network architecture needs to change to a ‘flexible per service’ model. The new 5G System of Systems will be built upon a strong foundation with an adaptive and programmable multi-service architecture. NFV and SDN are underlying principles for improving resource usage efficiency thus reducing the operator’s Total Cost of Ownership.
- The current LTE and EPC architecture is not able to efficiently serve the large variety of use cases; it is not able to provide low latency services, especially when they need full mobility support. In addition, the explosion of mobile broadband and massive deployment of MTC (aka IoT) will drive the overall complexity of the network and related cost, thus endangering a sustainable business.

- Rome wasn't built in a day: The road to this 5G network architecture transformation is not trivial and there are different market needs around the globe. Nokia believes that the new 5G radio and architecture should be introduced in phases:
  - The first phase solves imminent business needs to boost mobile broadband capacity at the same time leverage existing LTE deployments by tightly connecting 5G radio access to the 4G network via dual connectivity.
  - The second phase introduces an optimal architecture for supporting the diverse use cases like massive MTC, critical MTC and extreme MBB. Furthermore, a radio agnostic core network will be introduced to provide the long term convergence for radio and fixed access. The network slicing concept allows tailoring of the network resources according to the needs of the application and the capabilities of the device.

With the phased approach operators will be able to leverage existing deployments to provide higher data rates, better capacity in the near term and at the same time, introduce a future-proof network architecture to support new use cases and services in the longer term. Thus the 5G System of Systems, will give operators the right tools to harvest the new business potential.

References

[1] NGMN 5G White Paper

[2] Nokia Networks Network architecture for the 5G era White Paper

[3] 3GPP TS 23.401 "GPRS enhancements for E-UTRAN access"

Glossary of Abbreviations

3GPP	Third Generation Partnership Project
B2B	Business to Business
B2C	Business to Consumer
CAPEX	Capital expenditures
CDN	Content delivery network
D2D	Device to Device
DEM	Dynamic Experience Management
EPC	Evolved Packet Core
IMS	IP Multimedia System
IoT	Internet of Things
LTE	Long Term Evolution
LTE-M	Long Term Evolution for Machines
M2M	Machine-to-Machine
MAC	Medium Access Control
MBB	Mobile Broadband
MME	Mobility Management entity
MTC	Machine Type Communication
NaaS	Network as a Service
NAS	Non-Access Stratum
OPEX	Operating expenditure
MIMO	Massive-input Massive-output
NFV	Network Function Virtualization
NYU	New York University
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RAT	Radio Access Technology
RF	Radio Frequency
SAE	Service Architecture Evolution
SDN	Software defined networking
TCO	Total cost of ownership
UE	User Equipment
VoLTE	Voice over LTE

Nokia
Nokia Solutions and Networks Oy
P.O. Box 1
FI-02022
Finland

Visiting address:
Karaportti 3,
ESPOO,
Finland
Switchboard +358 71 400 4000

Product code C401-011929-WP-201510-1-EN