

5G Radio Access Network Architecture: Design Guidelines and Key Considerations

Patrick Marsch, Icaro Da Silva, Ömer Bulakci, Milos Tesanovic, Salah Eddine El Ayoubi, Thomas Rosowski, Alexandros Kaloxylou, and Mauro Boldi

The authors provide a comprehensive overview of the 5G RAN design guidelines, key design considerations, and functional innovations as identified and developed by key players in the field. They depict the air interface landscape that is envisioned for 5G, and elaborate on how this will likely be harmonized and integrated into an overall 5G RAN.

ABSTRACT

While there is clarity on the wide range of applications that are to be supported by 5G cellular communications, and standardization of 5G has now started in 3GPP, there is no conclusion yet on the detailed design of the overall 5G RAN. This article provides a comprehensive overview of the 5G RAN design guidelines, key design considerations, and functional innovations as identified and developed by key players in the field.¹ It depicts the air interface landscape that is envisioned for 5G, and elaborates on how this will likely be harmonized and integrated into an overall 5G RAN, in the form of concrete control and user plane design considerations and architectural enablers for network slicing, supporting independent business-driven logical networks on a common infrastructure. The article also explains key functional design considerations for the 5G RAN, highlighting the difference to legacy systems such as LTE-A and the implications of the overall RAN design.

INTRODUCTION

After several years of research on fifth generation (5G) wireless and mobile communications, there is broad consensus that 5G will not just be a “business-as-usual” evolution of 4G networks with new spectrum bands, higher spectral efficiencies, and higher peak throughput, but will also target new services and business models. The main 5G service types typically considered are extreme mobile broadband (xMBB, a.k.a. eMBB) with data rates up to several gigabits per second in some areas and reliable broadband access over large coverage areas, massive machine-type communications (mMTC) requiring wireless connectivity for, for example, millions of power-constrained sensors and actuators, and ultra-reliable MTC (uMTC, a.k.a. ultra-reliable and low-latency communications, URLLC) requiring end-to-end latencies of less than 5 ms and 99.999 percent reliability for, say, vehicle-to-everything (V2X) communication [1–3].

Investigations toward an overall 5G radio access network (RAN) architecture that can efficiently support these requirements are still ongoing. This article provides an overview of the

5G RAN design guidelines, key design considerations, and functional innovations identified and developed by key players in the field.¹ Key 5G RAN design requirements are listed, after which the envisioned 5G air interface (AI) landscape is described, including the latest considerations on how different air interface variants (AIVs) may be integrated into one overall AI. The article further captures overall system architecture considerations, such as the logical split between core network (CN) and RAN and related network interfaces, before venturing into key 5G functional design paradigms. Finally, conclusions are drawn.

KEY 5G RAN DESIGN REQUIREMENTS

Due to the diverse and extreme requirements of the main 5G service types, it is clear that the 5G RAN must be designed to operate over a wide range of spectrum bands with diverse characteristics, such as channel bandwidths and propagation conditions [1]. It must further be able to scale to extremes w.r.t. throughput, number of devices, connections, and so on, which is facilitated if the user plane (UP), related to application payload transmission, and control plane (CP), related to control functionality and signaling, are handled individually. For scalability also toward various possible deployments and an evolving application landscape, both 5G RAN and CN must be software configurable, meaning, for example, that it is configurable which logical and physical entities are traversed by CP and UP packets.

A common understanding is that the 5G RAN should allow integrating Long Term Evolution-Advanced (LTE-A) evolution and novel 5G radio on the RAN level, although integration need not always take place on this level.

The 5G RAN should further support more sophisticated mechanisms for traffic differentiation than LTE-A in order to fulfill diverse and more stringent quality of service (QoS) requirements, and facilitate the network slicing vision from Next Generation Mobile Networks (NGMN) [2], enabling independent operation of logical networks for different business cases on a shared physical infrastructure.

Another required feature distinctive from LTE-A is the native and efficient support of

¹ Note that throughout the article, any notion of “it is envisioned” or “it is proposed” refers to the views and proposals of partners of the METIS-II project: www.metis-ii.5g-ppp.eu/

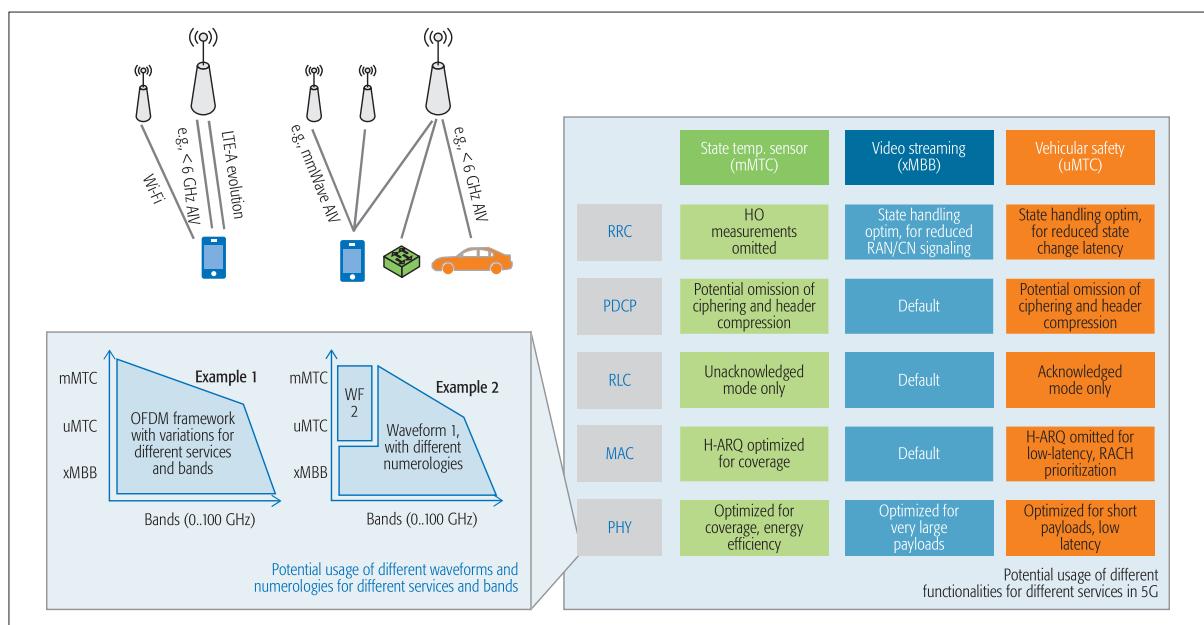


Figure 1. Overall 5G AI envisioned, composed of multiple AIVs with potentially different functionalities for different services and bands.

communication forms like multi-connectivity (e.g., concurrent communications of a device with multiple network nodes) and network-controlled device-to-device (D2D) communication, in the form of point-to-point, multicast, or broadcast communication. The 5G RAN should further support a wide range of physical deployments, from distributed base stations to centralized cloud-RAN deployments or distributed edge clouds. Different types of backhaul and fronthaul shall also be supported with steady performance degradation in case of reduced backhaul or fronthaul quality in terms of delay and capacity. Also, self-backhauling is seen as an important feature, where devices may also act as base stations and self-establish wireless backhaul links to suitable donor base stations.

Finally, the 5G RAN must be highly energy-efficient and future-proof, that is, enabling efficient introduction of new features and services and backward compatibility of devices in future releases.

AIR INTERFACE LANDSCAPE AND INTEGRATION INTO ONE 5G AIR INTERFACE

In order to simultaneously handle the extreme requirements of 5G use cases, an overall 5G AI is envisioned, shown in the top left of Fig. 1, which is composed of LTE-A evolution and novel AIVs tailored to different services, bands, and cell sizes. Also, non-cellular standards like Wi-Fi are considered as part of 5G, although they are not covered in this article. The overall 5G AI is expected to operate over a wide range of spectrum bands, where frequencies below 6 GHz are likely most suitable to support, for example, mMTC services where coverage is important, while spectrum above 6 GHz (up to 100 GHz) is essential to meet xMBB capacity demand. Three

authorization schemes or mixtures thereof are expected to coexist for 5G spectrum usages: primary user mode, licensed shared access (LSA), and unlicensed mode. Exclusive use of spectrum should remain the main and preferred solution, while shared use of spectrum may be a complement to increase spectrum availability, as summarized in [1].

Evolved LTE-A will likely play a pivotal role in the overall 5G AI, with tight interworking with novel AIVs [4], that is, support of inter-AIV mobility with very low interruption times and inter-AIV multi-connectivity. Note that this does not imply LTE assistance, but also stand-alone operation of novel AIVs or their usage as an anchor layer is considered.

Novel 5G AIVs are needed in particular to respond to stringent latency and reliability targets that LTE-A evolution cannot meet, or to enable communication at significantly higher frequencies than in existing systems. For this, some new AIVs are expected to use significantly shorter transmission time intervals (TTIs) and a wider bandwidth compared to LTE. It is also clear that there cannot be a “one-size-fits-all” solution for all novel AIVs. For example, an AIV tailored to lower carrier frequencies, large cell sizes, and high velocity will likely have a physical layer (PHY) designed to be most robust toward delay spread and Doppler spread, whereas an AIV for mmWave frequencies and short-distance communication with limited mobility might instead require robustness toward other impairments like phase noise. Key properties of novel 5G AIVs will be a flexible and scalable numerology (e.g., frequency-dependent TTI length) enabling user- and service-specific adaptations, flexible sub-band configurations, improved spectral and energy efficiency, support for flexible time-division duplex (TDD), reduced out-of-band (OOB)

For the CP, a single set of RRC specifications is envisioned among all novel AIVs. These might differ, however, for evolved LTE-A and novel AIVs in order to simplify the standardization process. Major changes envisioned in the RRC for novel AIVs are the need to support procedures relying on a beam-centric and lean design.

emissions, and support for non-orthogonal multiple access (NOMA) schemes [5].

A key challenge is the integration of different AIVs, including LTE-A evolution, into one overall 5G AI with minimal standardization and implementation complexity, while not sacrificing the performance of individual AIVs. In this respect, it is expected that there will be network functions (NFs) on different protocol stack layers that are tailored to different services (the right side of Fig. 1). For safety-critical vehicular communications, for example, one may need specific medium access control (MAC) functionalities like specialized hybrid automated repeat request (HARQ) for fast multicast communications and prioritized random access channels (RACHs). Together with service-specific higher-layer optimizations like the omission of header compression, the overall latency of emergency message transmission may be substantially reduced [6]. However, it appears desirable to design all NFs such that they can be parametrized from common specifications, and/or be selected from a common pool of atomic functions, for easier standardization and implementation.

Regarding PHY waveforms and numerology, it is envisioned that variations for different bands and services will be required (the bottom left of Fig. 1). It has not yet been concluded whether one would use a single waveform family such as orthogonal frequency-division multiplex (OFDM) parameterized to support all services and bands, or coexistence of different waveforms such as variants of OFDM and filter-bank multi-carrier (FBMC) [7]. However, one should strive for a large extent of harmonization, maximum reuse of signal processing blocks, and so on. While a large degree of NF harmonization among novel 5G AIVs can be enabled by design, the harmonization among novel AIVs and LTE-A evolution is more challenging, as this may pose too stringent limitations on 5G possibilities.

In conclusion, 5G will consist of a diverse set of AIVs, harmonized to a large but not complete extent, requiring novel architectural and functional design paradigms, as covered in the following sections.

OVERALL 5G SYSTEM ARCHITECTURE CONSIDERATIONS

CONSIDERATIONS ON CN, CN/RAN SPLIT, AND NETWORK INTERFACES

In alignment with NGMN [2], it is expected that the majority of CN and service-layer functions are deployed in 5G as virtual network functions (VNFs), running on virtual machines on standard servers, potentially on cloud computing infrastructures (i.e., data centers). Although not necessary, the function design will to some extent explore software-defined networking (SDN) principles such as UP/CP split, and allow for their flexible deployment in operators' networks depending on latency requirements, available transport, processing and storage capacity, and so on. Moreover, different services or network slices may utilize different CN and service-layer VNFs deployed at different network sites.

An important assumption, also considered

by the Third Generation Partnership Project (3GPP) [8], is a logical split between RAN, CN, and service layer functions. This appears beneficial because it allows independent evolution of RAN and CN functionality to accelerate the introduction of new technology, and enables access-agnostic CN functions (e.g., common UP processing). A downside is that cross-CN/RAN optimizations (e.g., collapsing of functionalities) are limited to single-vendor environments.

Regarding the exact logical split between RAN and CN, various changes are being proposed w.r.t. the evolved packet system (EPS) [9]. For example, paging is envisioned to be moved toward the RAN, and a new radio resource control (RRC) state has the potential to simplify the 5G CN, both aspects detailed later.

In previous generations, different radios like 3G and 4G were connected via different CNs and performed CN-based mobility via CN interfaces for handovers and location updates. For 5G, requirements related to latency, reliability, and short interruption times clearly render a tighter interworking between LTE-A evolution and novel AIVs necessary [4] (e.g., enabling RAN-based mobility and multi-connectivity). In this context, it appears beneficial to have a common CN/RAN interface for LTE-A evolution and novel AIVs, herein denoted as S1*, for:

- Fast establishment of multi-connectivity for any user equipment (UE) entering the system via LTE-A or novel AIVs, requiring neither CN/RAN nor non-access stratum (NAS) signaling
- Simplified UE implementation due to a common NAS layer for LTE-A and novel AIVs
- Common context handling enabling dynamic UE activity and roaming across LTE-A and novel AIVs without CN/RAN signaling involved
- Needing only one CN state, reducing the risk of losing state synchronization

One potential drawback of a common CN and CN/RAN interface is the need for a common evolution track for LTE-A and new AIVs, which may be complicated if the 5G CN should diverge strongly from the current design. Note that the envisioned S1* interface would inherently enable tight Wi-Fi integration as in LTE Release 13 via Xw interface.

Also, the existing X2 interface is expected to evolve (denoted herein X2*), in particular to support inter-node multi-connectivity, high-performance mobility, and RAN-based paging. The left side of Fig. 2 illustrates the envisioned logical architecture with mentioned interfaces.

USER PLANE AND CONTROL PLANE ARCHITECTURE

For the UP, a harmonized packet data convergence protocol (PDCP) for LTE-A evolution and novel AIVs can be achieved with a single set of functions captured in a common specification, potentially with different operation modes. This would facilitate the standardization process and the implementation of features such as handover and UP aggregation among AIVs. For aggregation among LTE-A evolution and novel AIVs, a bearer split should be supported where anchoring may occur on either side.

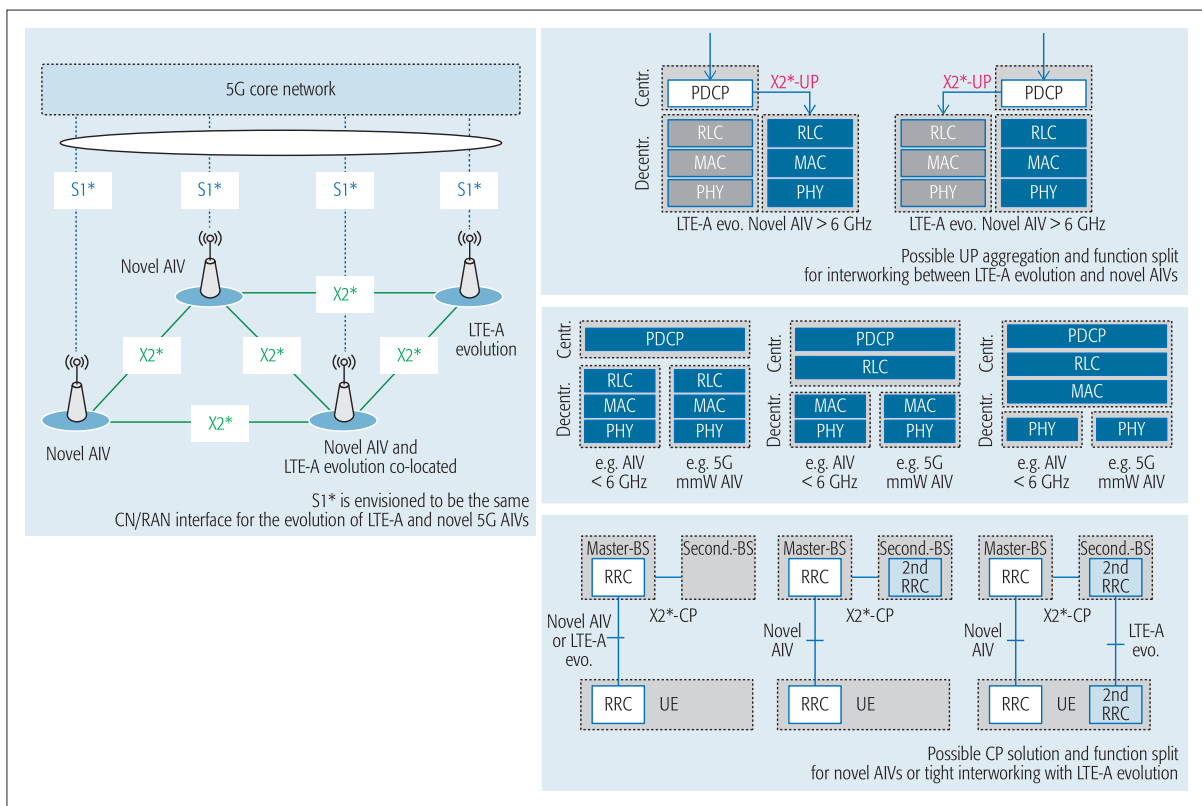


Figure 2. Logical architecture and network interfaces, options for multi-AIV integration, and function splits.

For the CP, a single set of RRC specifications is envisioned among all novel AIVs. These might differ, however, for evolved LTE-A and novel AIVs in order to simplify the standardization process. Major changes envisioned in the RRC for novel AIVs are the need to support procedures relying on a beam-centric and lean design.

For the CP support of multi-connectivity among novel AIVs, a single RRC connection is envisioned (as in LTE dual connectivity), since it accelerates the establishment of a secondary AIV and reduces UE complexity. For multi-connectivity among LTE-A evolution and novel AIVs, different options in Fig. 2 have been investigated, with the current assumption being to have a single RRC connection (i.e., a single state machine at the UE). The stated UP and CP architecture options are shown on the right side of Fig. 2.

DEPLOYMENT ARCHITECTURE, PLACEMENT OF LOGICAL FUNCTIONS, AND RECONFIGURATION

As mentioned before, the 5G RAN should support a wide range of physical deployments and maximally leverage centralization, while also supporting distributed base stations and non-ideal backhaul. A key enabler for this is the implementation of some radio functions as VNFs, as mentioned before in the CN context, allowing flexibly shifting these toward or away from the radio edge depending on the physical architecture and specific application requirements. Preliminary analyses [10] concluded that functions that are time-*asynchronous* to the

radio interface (in LTE these are PDCP and RRC functions related to measurement control and reporting, handover preparation and execution, dual connectivity, random access, RRC state transition, etc.) are mostly suitable to be implemented as VNFs and possibly centralized, as they typically require low data rates on their interfaces, scaling with the number of users and not the overall traffic. Further, these functions can typically cope with larger latency (e.g., tens of milliseconds in LTE).

However, time-*synchronous* functions (in LTE these are PHY, MAC, and radio link control [RLC] functions, e.g., scheduling, link adaptation, power control, interference coordination) typically require high data rates on their interfaces, scaling with the traffic, signal bandwidth, and number of antennas, and benefit from hardware acceleration. The potential for centralization is here most pronounced in deployments with low-latency and high-bandwidth backhaul/fronthaul due to timing and real-time processing requirements.

A key consideration for 5G is to design RAN functions such that strict timing relations between the protocol layers are avoided, and hence multiple possible function splits among physical entities are possible. The right side of Fig. 2 shows some possible function splits, including the aforementioned split between asynchronous and synchronous functionality.

The 5G RAN is envisioned to enable a fast reconfiguration of the mix of supported AIVs and frequency bands, the reconfiguration of the

Beamforming will play an essential role in 5G, in particular in the context of bands beyond 6 GHz, where a larger path loss has to be overcome, but where shorter wavelengths also permit a larger number of antenna elements at reasonable form factor. Different from LTE-A, where only data channels use beamforming, higher bands will also require beamforming of control channels.

related NFs as such, and the mapping of NFs to physical architecture. This reconfiguration functionality is expected to act on the basis of various input parameters, such as available resources (spectrum and hardware), traffic demand, UE capabilities (supported AIVs, frequency bands, etc.), and requested services (e.g., bandwidth and quality of service [QoS]), and could exploit collaboration among network nodes.

RAN ARCHITECTURE SUPPORT FOR NETWORK SLICING

Flexibility and configurability will be key RAN characteristics to support the stated diverse services and related requirements in one common network infrastructure. This may be realized by a protocol architecture supporting a service-specific selection of NFs and service-tailored optimizations, as elaborated earlier. Going beyond the support of diverse services as such, network slicing [2] refers to the operation of multiple independent logical networks for different business cases on a shared physical infrastructure. For instance, a dedicated network slice could be configured for a particular V2X business case, involving service level agreements (SLAs) among stakeholders and a slice-specific selection and configuration of CN, transport network, and RAN functions. It is envisioned that [6]:

- Even if network slices are seen as separate logical networks, efficient reuse of resources like radio spectrum, infrastructure, transport network, and NFs among the slices is essential.
- Network slices (or an abstraction thereof, e.g., groups of service flows) need to be visible to the RAN, such that NFs can take into account overall slice-specific metrics or constraints (e.g., the constraint that all services belonging to a slice may jointly only occupy a certain amount of radio resources).
- Means for slice isolation and protection are needed (e.g., to guarantee that events in one slice cannot negatively impact another slice).
- Performance monitoring solutions need to be aggregated per slice to verify the fulfillment of SLAs and/or properly operate the different businesses associated with different slices.
- Configuration management, self-organizing networks (SON), and so on could be slice-tailored. Additionally, certain functions could be turned on/off and/or possibly configured individually for different slices. This decision should be made before the deployment of a slice; however, the fine-tuning of the configuration of the functions shall be possible during the lifetime of the slice.

One open question related to network slicing is how the system could tell to which slice a newly appearing device or service should belong. This could be solved via network implementation where UE identities are matched with the slice to which they belong. Alternatively, a slice-specific identity could be either transmitted over system information to support the UE to access the proper slice or a slice-aware authentication/attach where the UE has some sort of slice-specific identity.

KEY FUNCTIONAL DESIGN PARADIGMS FOR THE 5G RAN

Beyond the overall system architecture, the envisioned service and band diversity suggests various paradigm changes in the functional design of the 5G RAN, as explained in the following.

BEAM-CENTRIC DESIGN

Beamforming will play an essential role in 5G, in particular in the context of bands beyond 6 GHz, where a larger path loss has to be overcome, but where shorter wavelengths also permit a larger number of antenna elements at reasonable form factor. Different from LTE-A, where only data channels use beamforming, higher bands will also require beamforming of control channels. For analog/hybrid antenna architectures, beam sweeping procedures are proposed for synchronization and reference signals where transmissions do not occur simultaneously over all directions. For initial access, a beam-based system increases inefficiency of broadcasted information, posing a severe issue given the size of the system information in LTE-A. Compression mechanisms are being proposed that minimize the information transmitted in common channels (possibly not beamformed), using dedicated channels for the remaining information when UEs are connected. Mobility reference signals will also require beamforming, and the UE should be able to detect different beams. In addition, due to the volatility of links in higher frequencies, faster and robust beam switching mechanisms should be supported including the UE directly using a pre-configured candidate beam. Clearly, a beam-centric design, which is necessary for higher bands, but may for harmonization reasons be generally introduced in 5G, has a large implication on the overall RAN design [11].

DEDICATED AND SELF-CONTAINED TRANSMISSIONS, AND NEW SCHEMES TO DISTRIBUTE SYSTEM INFORMATION

In LTE-A, mobility and system access procedures rely on “always-on” signals such as synchronization sequences, cell-specific reference signals, and system information. For a lean and future-proof design, active mode mobility should rely, whenever possible, on more dynamically configurable reference signals possibly sparsely transmitted in time and in specific frequency resources. These could be either periodic or configured on demand based on mobility scenarios (highways, semi-static environments, etc.). Besides minimization of broadcast signals, as depicted in Fig. 3a, one should strive to avoid transmitting, for example, reference signals over the entire bandwidth, but instead use self-contained transmissions as shown in Fig. 3b. Here, reference signals used for the estimation of data channels are transmitted jointly with the payload, minimizing the overhead and interference, and are better suited for beam-based transmission and more future-proof (e.g., facilitating the definition of new physical channels when needed) [12].

In particular, in the context of the co-usage of lower and higher frequency bands, new ways to distribute common signals are envisioned, for instance, letting only some nodes, frequency carriers, or AIVs transmit system information or

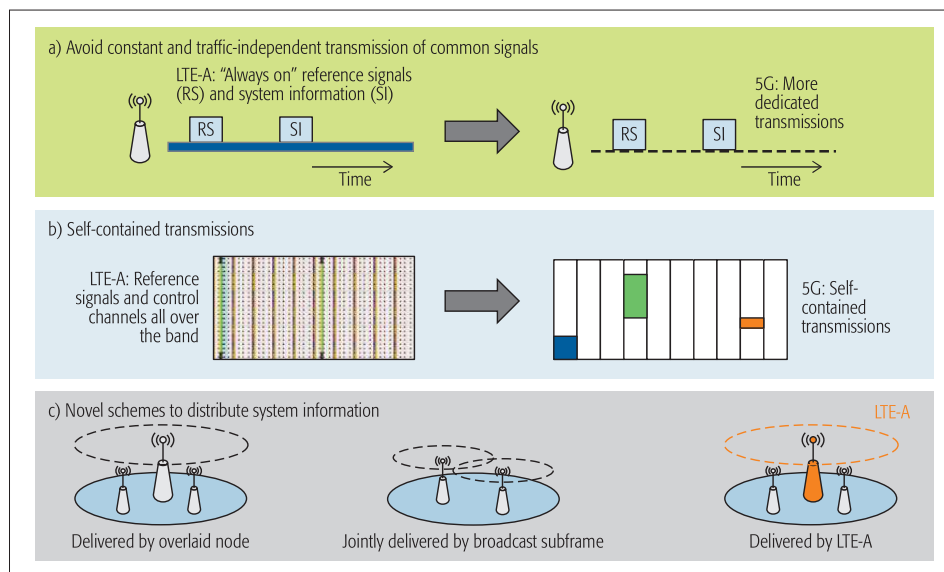


Figure 3. Dedicated and self-contained transmissions, and means to distribute system information in 5G.

parts of it (Fig. 3c). These solutions exploit the fact that multiple carriers and nodes might use exactly the same system information configuration.

SERVICE PRIORITIZATION AT INITIAL ACCESS

In LTE-A, service prioritization is possible through access class barring, where resources are split among services with significantly diverse QoS requirements. A proposed paradigm change in 5G is to provide service differentiation already on the PHY/MAC level. As one specific example, random access requests associated with delay-sensitive services could apply a combination of preamble signatures at a given random access time slot, enabling high-priority service requests to have very small collision rates (up to 100 times less than state-of-the-art access class barring schemes [13]), since they could easily be identified at the receiver side. Without service differentiation at the initial access phase, uncritical services may impair the successful initial access of mission-critical services.

HIERARCHICAL AND DYNAMIC TRAFFIC STEERING AND RESOURCE MANAGEMENT IN 5G

A key challenge in 5G is how to dynamically assign the foreseen wide service range to the different spectrum bands, usage types, related AIVs, and the radio resources therein, also involving novel communication forms like unicast or multicast network controlled D2D. This challenge will further be pronounced through more dynamic network topologies, for instance, resulting from a flexible activation and deactivation of access nodes, antennas and remote radio heads (RRHs), moving cells, and nomadic nodes. Furthermore, interference interdependencies between network nodes will increase, for instance, in dense deployments with TDD and a flexible usage of uplink (UL) and downlink (DL), D2D, or in-band self-backhauling. Consequently, key design principles related to traffic

steering and resource management need reconsideration [13].

A proposed paradigm change in 5G is an extended notion of the term *resource*, considering not only classical radio resources in time and frequency, but also soft network capabilities like processing power, memory/storage capacity and power budget. Also, licensed spectrum resources should be natively complemented by shared (e.g., LSA) or unlicensed resources, where the duration and extent of spectrum sharing may depend on the uncontrolled source of interferers and dynamicity in topology changes.

In existing systems, the assignment of services to radio access technologies (RATs, e.g. 3G and 4G) and cells takes place via handover (i.e., RRC-level mechanisms). In 5G, in particular considering more stringent QoS requirements and the relative unpredictability of radio links in higher bands, traffic steering will need to be performed in a more agile way and on a faster timescale.

One specific proposal is a hierarchical traffic steering and resource management approach in the RAN, as illustrated in Fig. 4. Here, the "outer loop" (access network-outer, AN-O) receives the QoS policies from the CN and maintains a cross-AIV 5G RAN view, based on real-time feedback (per-TTI or periodically over a few TTIs) from the AIVs in the "inner loop" (access network-inner, AN-I). This allows the AN-O to do fast traffic re-routing among the AIVs, and also dynamically (de-)activate AIVs and place UEs into suitable multi-AIV multi-connectivity constellations, while the actual TTI-level resource management happens in the AN-I. The exact implementation of the scheme depends on the chosen function split in the RAN. If, for instance, RRC, PDCP, and RLC, are centralized in a cloud RAN, and MAC and PHY are performed in radio units, the former group would naturally constitute the AN-O and the latter the AN-I, as depicted on the right side of Fig. 4. The

A key challenge in 5G is how to dynamically assign the foreseen wide service range to the different spectrum bands, usage types, related AIVs and the radio resources therein, also involving novel communication forms like uni-cast or multi-cast network controlled D2D. This challenge will be further pronounced through more dynamic network topologies.

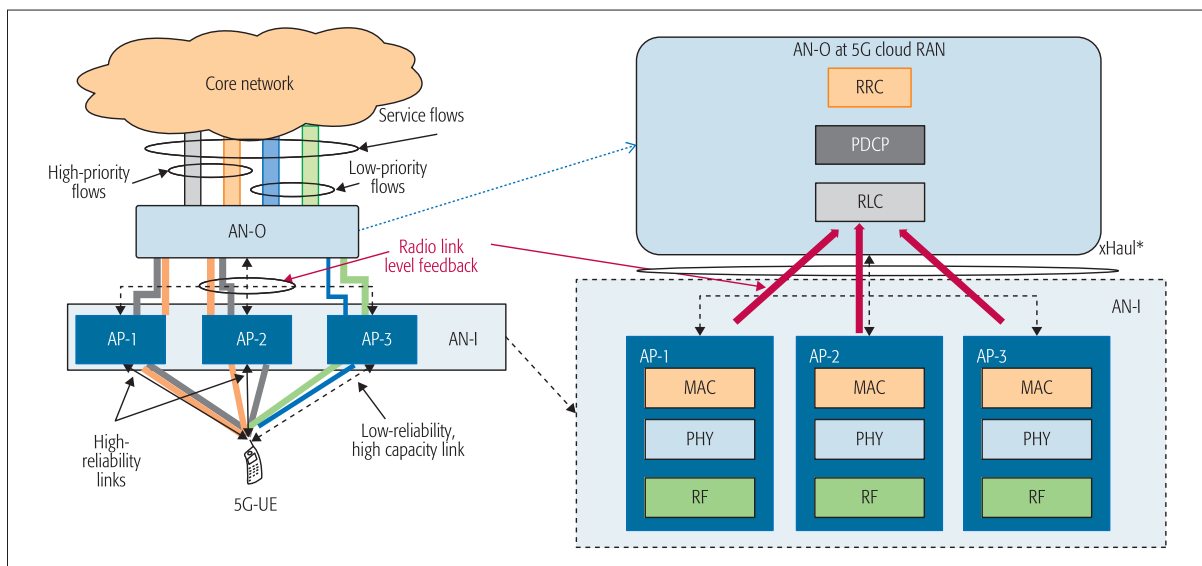


Figure 4. An example scheme for hierarchical traffic steering and resource management in its abstract form (left), and in the context of a RAN function split (right). The function split shown on the right illustrates one specific example, while different split options depending on xHaul* characteristics are possible.

interface between AN-O and AN-I, here denoted as xHaul*, would then have to encompass new RAN measurement information elements for the real-time AIV feedback to the AN-O, and control information elements for the steering of the AN-I resource management schemes by the AN-O. It is still under research to which extent these information elements should be standardized and may possibly also require changes to the Uu interface.

A NOVEL RRC STATE MODEL IN 5G

Moving toward the Internet of Things (IoT), there will be significantly more battery-powered UEs (e.g., sensors, baggage tags, wearable devices) in 5G. Therefore, battery efficiency and duration will be essential, especially for devices with limited accessibility (e.g., in remote locations or restricted areas). However, fast first packet transmission (UL or DL) may be even more important than today. This trade-off between device power efficiency and fast accessibility is often called the “UE sleeping problem.”

For 5G, a novel RRC state model is proposed to address the problem, relying on a novel state, RRC Connected Inactive, in addition to RRC Connected and RRC Idle. This novel state explores the principle of “not discarding previously exchanged information” for inactive UEs, meaning that UEs in the new state still keep parts of the RAN context (e.g., related to security, UE capability information). In addition, signaling is reduced by allowing the UE to move around within a pre-configured area without notifying the network, even among LTE-A evolution and novel AIVs. The new state is envisioned to be highly configurable with a wide range of discontinuous reception (DRX) cycles (from milliseconds to hours) and service-tailored state transitions.

Transitions from RRC Idle to RRC Connected are expected to occur mainly when a UE first

attaches to the network, or as a fallback when devices and/or the network cannot use the previously stored RAN context. However, transitions from RRC Connected Inactive to RRC Connected are expected to occur often and are hence optimized to be fast and lightweight in terms of signaling. This is achieved by keeping the CN/RAN connection alive during inactivity periods and reducing the amount of RRC signaling needed to resume an existing inactive connection via the usage of a RAN context ID — an option inspired by the suspend/resume procedure considered for idle state UEs in LTE [14]. Thanks to the common CN/RAN interface, the UE is able to perform this optimized state transition even when roaming among LTE-A evolution and novel AIVs. The proposed novel state model and expected benefits w.r.t. protocol overhead and CP latency reduction are shown in Fig. 5, with more details and a discussion on similarities and differences of the proposed new RRC state to the URA-PCH state in high-speed packet access (HSPA) given in [15].

RAN-BASED PAGING

Due to the expected massive number of devices and denser deployments in 5G, paging may significantly increase the load on the AI and the CN/RAN interface, thus requiring new solutions for efficient paging and UE location tracking. One proposed option is hierarchical location tracking where the CN tracks UEs in RRC Connected Inactive on the level of groups of RAN locations, whereas the RAN tracks these at cell-level granularity. This can be seen as a combination of the hierarchical location concepts in high-speed packet access (HSPA) with the leaner state model (compared to HSPA) described before [15]. This could involve a lightweight signaling procedure terminated in the RAN, using a security handling mechanism based on retaining and updating the security context from the last attach procedure. The proposed hierarchical

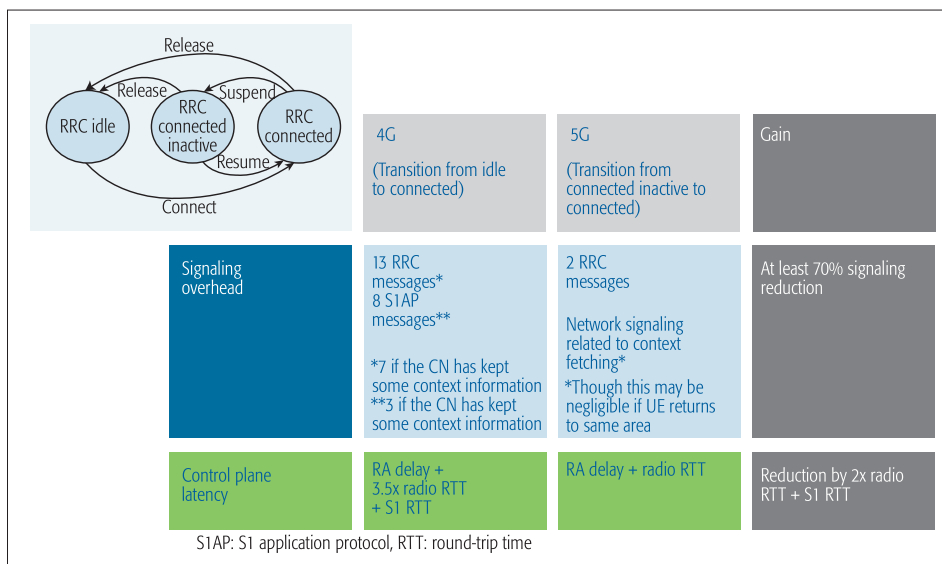


Figure 5. The proposed novel RRC state model and its expected benefits.

5G functional design paradigm	Key benefits	Key difference to LTE-A evolution	Implication on overall RAN design
Beam-centric design	Better coverage, capacity and data rates in higher bands	Narrow beams possibly swept instead of notion of omnidirectional cells	All control signals beamformed. All mobility and initial access procedures need native beam-centric design.
Dedicated and self-contained transmissions, new ways of distributing system information	Energy efficiency and future-proofness, potentially also improved C-plane scalability	Reference signals not always on, not full band, not all subframes	Significantly more configurable reference signals and mobility procedures
Service prioritization at initial access	Service differentiation already at first access, lower latency, and 100x reduction of collision probability for mission-critical services [13]	Different levels of service prioritization for diverse sets of delay requirements	New MAC procedures required for RACH to enable service prioritization. Signaling to higher layers needed.
Hierarchical and dynamic traffic steering and resource management	Faster and more agile resource management, around 20% reduction in mean packet delivery delay [13], higher overall reliability. Essential for uMTC.	Traffic steering performed on comparatively lower protocol stack layer. Dynamic traffic steering instead of handover.	New control information elements and steering options between protocol layers needed. Possible impact on Uu interface.
A novel RRC state model	Reduced UE power dissemination and C-plane latency. Reduction of CN/RAN signaling by at least 70% [15]. Especially suitable for bursty connectivity and massive access.	UEs are always connected from a CN perspective. Significantly larger possibilities for service-specific configuration.	Context fetching needs to be specified and supported. Novel mobility procedures for new state need to be defined.
RAN-based paging	CN/RAN signaling reduced by more than 85%, reduced C-plane latency [13].	In LTE, paging is a CN function, which is now moved into the RAN	Entire re-design of paging functionality, signaling etc. Change of usage of CN/RAN interface.

Table 1. Summary of the key functional design paradigms considered.

location tracking approach would imply moving some paging and mobility related functionalities from CN to RAN, hence changing the CN/RAN split compared to EPS.

Table 1 summarizes the key functional design paradigms explained before, their key benefits, difference from LTE-A, and their implication on the overall 5G RAN design.

CONCLUSIONS

This article has provided a consolidated view of the likely overall 5G RAN architecture, functions, and interfaces, enabling integration of

LTE-A evolution and multiple novel AIVs in 5G. Besides providing details on the likely logical architecture and its mapping to physical architecture as well as means to support network slicing, the article has ventured into key functional paradigm changes that are proposed (e.g., related to a beam-centric design, lower-layer service prioritization, traffic steering in 5G, and a novel RRC state model), emphasizing the key differences from LTE-A and the main RAN design implications. It has to be noted that the topics covered in this article are still under research and subject to finalization in the next months.

Due to the expected massive number of devices and denser deployments in 5G, paging may significantly increase the load on the AI and the CN/RAN interface, thus requiring new solutions for efficient paging and UE location tracking.

ACKNOWLEDGMENT

This work has been performed in the framework of the 5G PPP project METIS-II co-funded by the European Union. The views expressed are those of the authors and do not necessarily represent the project. The authors would like to thank Gerd Zimmermann, Panagiotis Spapis, and Athul Prasad for their contributions and suggestions.

REFERENCES

- [1] ICT-671680 METIS-II, Deliv. 1.1, "Refined Scenarios and Requirements, Consolidated Test Cases, and Qualitative Techno-Economic Assessment," Jan. 2016.
- [2] NGMN Alliance, "NGMN 5G White paper," Feb. 2015.
- [3] ITU-R Working Party WP 5D: Draft New Rec. ITU-R M.2083-0 (09/2015), "IMT Vision – Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond," Sept. 2015.
- [4] 3GPP TR 38.913, "Study on Scenarios and Requirements for Next Generation Access Technologies," Mar. 2016.
- [5] L. Dai *et al.*, "Non-Orthogonal Multiple Access for 5G: Solutions, Challenges, Opportunities, and Future Research Trends," *IEEE Commun. Mag.*, vol. 53, no. 9, Sept. 2015, pp. 74–81.
- [6] I. Da Silva, G. Mildh *et al.*, "On the Impact of Network Slicing on 5G Radio Access Networks," *Euro. Conf. Networks and Commun.*, Athens, Greece, June 2016.
- [7] M. Tسانovic *et al.*, "Towards a Flexible Harmonised 5G Air Interface with Multi-Service, Multi-Connectivity Support," ETSI Wksp. Future Radio Technologies: Air Interfaces, Jan. 2016.
- [8] 3GPP S2-153651, "Study on Architecture for Next Generation System," Oct. 2015.
- [9] 3GPP TS 36.300, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall Description; Stage 2," Nov. 2015.
- [10] ICT-317669 METIS, Deliv. 6.4, "Final Report on Architecture," Jan. 2015.
- [11] F. Athley *et al.*, "Providing Extreme Mobile Broadband Using Higher Frequency Bands, Beamforming and Carrier Aggregation," *IEEE Int'l. Symp. Personal, Indoor and Mobile Radio Commun.*, Aug. 2015, pp. 1370–74.
- [12] P. Frenger, M. Olsson, and E. Eriksson, "A Clean Slate Radio Network Designed for Maximum Energy Performance," *IEEE Int'l. Symp. Personal, Indoor and Mobile Radio Communications (PIMRC)*, Sept. 2014, pp. 1300–04.
- [13] ICT-671680 METIS-II, Deliverable 2.2, "Draft Overall 5G RAN Design," June 2016.
- [14] 3GPP TR 23.720, Technical Specification Group Services and Systems Aspects, "Architecture Enhancements for Cellular Internet of Things (Rel. 13)," Nov. 2015.
- [15] I. Da Silva *et al.*, "A Novel State Model for 5G Radio Access Networks," *IEEE ICC Wksp.*, May 2016, pp. 632–37.

BIOGRAPHIES

PATRICK MARSCH (patrick.marsch@nokia-bell-labs.com) received his Dipl.-Ing. and Dr.-Ing. degrees from Technische Universität Dresden, Germany, in 2004 and 2010, respectively. After leading a research group at TU Dresden, Germany, since 2011 he has been heading a research department within Nokia Bell Labs (formerly Nokia Siemens Networks), Wrocław, Poland. He has published 60+ journal or conference papers, co-edited 2 books, received 4 best paper awards, and received the Philipp Reis Prize for pioneering work in the field of coordinated multipoint. He has been General Chair of various IEEE workshops, and has initiated the ULTRA² series of workshops on ultra-low latency and ultra-high reliability in wireless communications. He is the Technical Manager of the 5G PPP project METIS-II.

ICARO DA SILVA is a senior researcher working at Ericsson, Stockholm, Sweden, since 2010. He is currently working as a 3GPP delegate driving technical areas such as 5G mobility and initial access. He was a Work Package Leader in the 5G-PPP project METIS-II responsible for the Overall Control Plane Design for the 5G architecture. He is also a co-author of the chapter about 5G architecture in the book *5G Mobile and Wireless Communications Technology* and

an author of many conference/journal papers in various areas from mobile networks to signal processing and physical layer. He is also an active inventor in the area of mobile networks, architecture, and network management with more than 100 patents. He received his M.Sc. from the Universidade Federal do Ceará, Brazil, in 2007 and since then he has been involved in R&D projects with Ericsson.

ÖMER BULAKCI received his B.Sc. degree in electrical and electronics engineering from the Middle East Technical University in 2006, his M.Sc. degree in communications engineering from the Technical University of Munich in 2008, and his doctoral degree in communications engineering from Aalto University in 2013. From 2009 to 2012, he conducted his research activities with emphasis on relay networks and standardization of relaying at Nokia Siemens Networks, Germany. Since October 2012, he has been conducting research toward 5G and contributing to EU flagship projects METIS and METIS II at Huawei Technologies ERC, Germany. He is an author of 50+ publications and is an inventor of 10+ patent applications. He is currently leading the agile resource management framework in METIS II.

MILOŠ TESANOVIC [SM] is principal 5G researcher at Samsung Electronics R&D Institute UK, focused on coordinating collaborative European R&D activities and conducting research into next-generation mobile systems. He is currently leading the 5G-PPP/H2020 METIS-II project's Work Package on 5G air interface design. He obtained his Ph.D. in 2008 from the University of Bristol, United Kingdom, for his work on robust wireless transmission of video over MIMO systems, and his graduate engineering (Dipl. Ing.) degree in electronics, telecommunications, and control in 2004 from the University of Belgrade, Serbia. He has filed more than 35 patents/patent applications, published over 20 refereed journal and international conference papers, and made active contributions to PHY and MAC standardization of LTE and LTE-A.

SALAH EDDINE EL AYOUBI received his Ph.D. and habilitation degrees from University Paris 6, France, in 2004 and 2009, respectively. Since 2004, he has been working at Orange Labs as a senior research engineer. His research interests include design and performance evaluation of mobile networks, and his current focus is on 5G RAN design. He is active in several EU research projects including 5G PPP METIS-II, where he acts as a Work Package Leader, and 5G-PPP FANTASTIC-5G project where he acts as Technical Coordinator.

THOMAS ROSOWSKI received his Diploma degree in electrical engineering from the University for Applied Science in Bingen, Germany, in 1986. He started his career in the area of broadcasting, got experience in international mobile network standardization, also holding posts as Chair and Secretary of Working Groups in ETNO and ETSI. Since 2009 he has been with Deutsche Telekom Laboratories where his areas of work include wireless and mobile technologies and networks, and radio spectrum management. In addition to company internal projects, he participated in the EU funded research projects E3 and METIS, and is currently involved in METIS-II. He is co-author of a number of publications on radio spectrum regulation and mobile network architectures.

ALEXANDROS KALOXYLOS [SM] received his B.Sc. from the University of Crete in 1992, his M.Phil. from Heriot-Watt University in 1994, and his Ph.D. from the University of Athens in 1999. He has participated in numerous EU projects and has published over 110 papers in the area of mobile communications since 1994. He is an Editorial Board member of several journals and a TPC member for numerous conferences. He is an assistant professor at the Department of Informatics and Telecommunications of the University of Peloponnese in Greece. Since June 2014, he is a principal researcher at Huawei's German Research Center in Munich, where he is currently the head of Radio Access Network Department.

MAURO BOLDI graduated at Politecnico di Torino, Italy, in 1998. He joined the Telecom Italia research centre (CSELT at that time) in the same year, where he has been involved in radio access network topics, with activities on antenna and propagation, radio over fiber solutions, collaborative networks, and energy efficiency. He is an author of patents and papers on the same topics, and he has been involved in many European projects, including Winner (where he led the activity on CoMP), Artist4G, Earth, and Metis/Metis-II. He is also a co-author of books on LTE-Advanced and 5G after the projects Winner and METIS. He is Telecom Italia delegate in the ETSI EE Technical Committee, where he plays the role of Rapporteur of the European Specification on Energy Efficiency for Radio Access Networks and Vice-Chairman of the EEPs group.