

5G NETWORK ARCHITECTURE

The 20th December 2017 will be remembered as an important day in telecommunications history as, on this day, during a meeting in Lisbon, Portugal, 3GPP (3rd Generation Partnership Project) successfully completed the first implementable 5G NR specification. NR (New Radio) is the term used to describe the 5G air interface and radio access network. This is the first phase of delivering a complete 5G end-to-end network based on the architecture presented in this article.

**ANDY
SUTTON**

Scalable and
optimised 5G
service delivery

The first mobile implementation of 5G is designed to work in Non-Standalone (NSA) mode to support the enhanced Mobile Broadband (eMBB) use case. In NSA mode the connection is anchored in Long Term Evolution (LTE) (3GPP 4G technology) with 5G NR carriers being used to increase data rate and reduce latency.

5G is often referred to as the next generation of mobile communications technology but the potential is more significant than this. 5G will likely become the future of communications, supporting fixed and mobile access. In addition to eMBB, 5G will support Ultra-Reliable and Low Latency Communications (URLLC), also referred to as Mission Critical Communications, and massive Machine Type Communications (mMTC) – an evolution of IoT – along with Fixed and Mobile Convergence. Although the diverse requirements of eMBB, URLLC and mMTC will not be supported from day-one, a flexible approach to the design of NR has been necessary to ensure 5G standards will evolve to meet all requirements. This approach has resulted in a NR design with scalable numerology (numerology refers to waveform parametrisation, e.g. cyclic prefix and subcarrier spacing in Orthogonal Frequency Division Multiplexing (OFDM)), numerology multiplexing and implementation of Time Division Duplex (TDD). TDD is better suited to data-centric services in which the downlink (the connection from network to user) will carry significantly more data traffic than the uplink (connection from user to network) in the vast majority of use cases. TDD will be the most common implementation across the majority of initial 5G frequency bands although it should be noted that Frequency Division Duplex (FDD) operation is also supported.

The December 2017 release of 5G NR does not include a 5G Next Generation Core (NGC) network but rather relies on an evolution of the existing 4G Evolved Packet Core (EPC) often referred to as EPC+. This means that a 5G-capable device will be

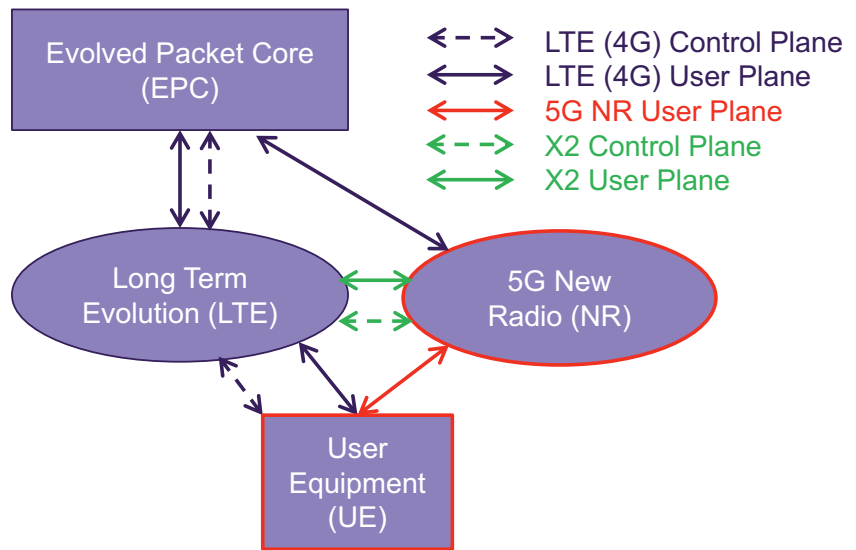


Figure 1: Option 3x 5G non-standalone network architecture

connected to an enhanced 3GPP Release 15 4G radio for control plane and 4G and/or 5G radio for user plane traffic flows. This concept is illustrated in Figure 1 noting that, in addition to the 3x architecture illustrated, there are other approaches to the connectivity between LTE and 5G NR to the Release 15 EPC.

Spectrum

Ofcom and the European Radio Spectrum Policy Group have identified three pioneer frequency bands for the introduction of 5G services in Europe. These bands are listed below along with the amounts of spectrum to be auctioned in each band for future 5G use in the UK:

- 700MHz with 2 x 30MHz (FDD) + 20MHz centre gap (supplementary downlink) to provide a wide-area coverage layer - spectrum auction expected during 2019.
- 3.4 to 3.8GHz with 150MHz of spectrum in 3.4 to 3.6GHz band and 116MHz of spectrum in 3.6 to 3.8GHz spectrum band (TDD) to provide a large amount of contiguous spectrum for high data rates and low-latency services, and also a capacity solution in congested areas – to be auctioned as two blocks, firstly the 3.4 to 3.6GHz band will be auctioned in 2018 with 3.6 to 3.8GHz to be auctioned during 2019.

- 24.25 to 27.5GHz (referred to as 26GHz band) with 3.25GHz of available spectrum to provide extremely high-speed data services and very low latency at short distances along with addressing future massive area capacity density requirements – to be auctioned in the future (no date set as of yet).

Spectrum in the frequency band of 3.4 to 3.6GHz will be the first new spectrum available in the UK for 5G use. This is known as band 42 in LTE although in 5G terminology it has been combined with LTE band 43 (3.6 to 3.8GHz) to form 5G band n78. Band n78 covers the 5G TDD spectrum range of 3.3 to 3.8GHz. Note that while designated, LTE bands 42 and 43 are not actually deployed in Europe.

The 3.4 to 3.6GHz band has a higher propagation loss than existing cellular frequency bands. During the early days of 5G rollout, this band will not necessarily offer contiguous coverage for both downlink and uplink communications and therefore the wider coverage of LTE, typically at 800MHz or 1800MHz, will support the control plane and in some scenarios, the user plane. One advantage of 5G is the adoption of massive Multiple Input, Multiple Output (MIMO) technology, an evolution of the MIMO technology we've seen in LTE but at a much larger scale. This

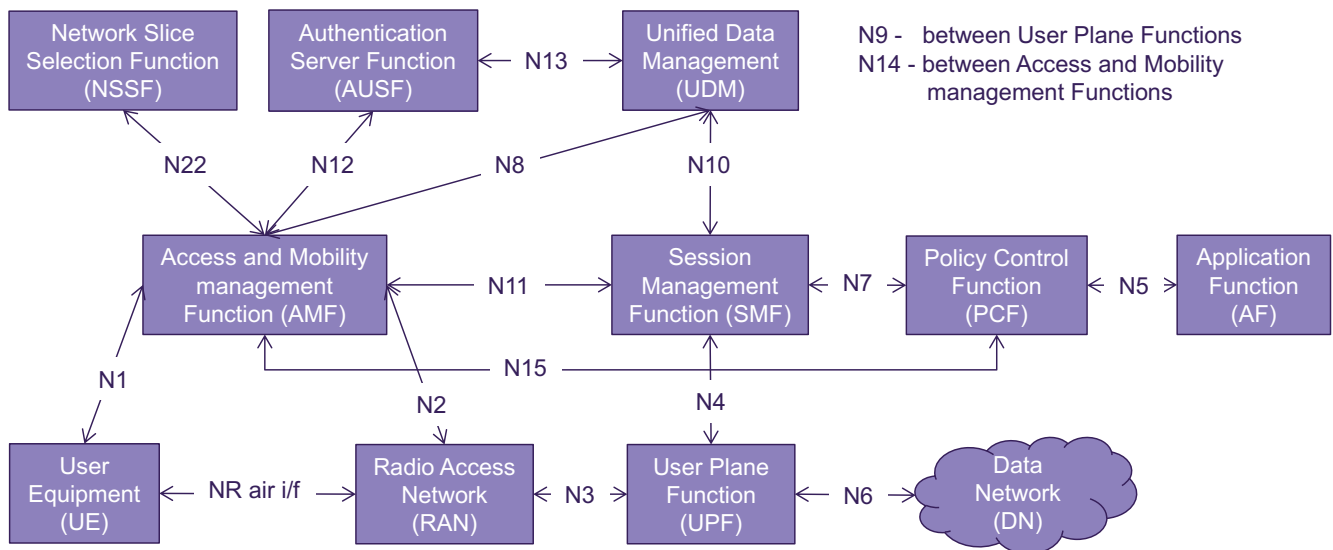


Figure 2: 5G network architecture

increased scale is realised by supporting 64 radio transceivers and antennas within the antenna module in such a way that beam-forming can be implemented to enhance the coverage by increased directional antenna gain. There is widespread global alignment behind the 3.4 to 3.8GHz band with ongoing discussions about extending this band in the future to 4.2GHz.

The higher frequency bands are not as well aligned; Japan, South Korea and the USA are favouring the 28GHz band rather than the 26GHz band which will be supported in Europe. There is a 1 GHz overlap between these two bands and it is anticipated that both will be fully supported by the 5G ecosystem.

While the main focus of 5G spectrum discussions are currently on new spectrum, any existing cellular frequency bands can and most likely will be re-farmed to 5G NR in the fullness of time.

3GPP network architecture

The remainder of this paper will focus on the complete 5G end-to-end network architecture which is the combination of 5G NR and NGC. 3GPP will complete standardisation of a 5G network architecture by June 2018 with Release 15 (phase 2) which supports subscriber data management, control plane functions

and user plane functions. Since the early days of Global System for Mobile Communications (GSM) and then General Packet Radio Service (GPRS) we’ve been familiar with logical representations of mobile network architectures. These diagrams take the form of functional blocks and the interfaces between them, officially known as reference points. Figure 2 presents this view of the 3GPP 5G network, referred to as “reference point representation”.

The reference points or interfaces, which will be known as interfaces for the remainder of this paper, start with the letter ‘N’. Originally these were designated ‘NG’ for next generation, however recently the term has been shortened to simply read ‘N’. The functional blocks are split between control plane and user plane functions with the control plane further split between subscriber management functions and control plane functions.

The subscriber management functions consist of the Authentication Server Function and Unified Data Management while the control plane function consists of a core Access and Mobility management Function, a Session Management Function, Policy Control Function, Application Function and Network Slice Selection Function (NSSF). The NSSF is responsible

for selecting which core network instance is to accommodate the service request from a User Equipment (UE) by taking into account the UE’s subscription and any specific parameters. The user plane functions start with the UE which may be a smartphone or a new form factor terminal, possibly fixed rather than mobile. This connects via the Radio Access Network (RAN) to the User Plane Function (UPF) and on to a Data Network (DN). The DN may be the Internet, a corporate Intranet or an internal services function within the mobile network operator’s core (including content distribution networks).

The NR air interface downlink waveform is Cyclic Prefix-Orthogonal Frequency Division Multiplex (CP-OFDM) access while the uplink can be either CP-OFDM or Discrete Fourier Transform-spread-Orthogonal Frequency Division Multiple access, the uplink mechanism being selected by the network based on use case. The UE connects to the RAN via the air interface which also carries the N1 interface which, in previous iterations of 3GPP technologies, has been known as the non-access stratum. This is a peer-to-peer control plane communication between the UE and core network.

The N3 interface is what is commonly known as mobile backhaul between the

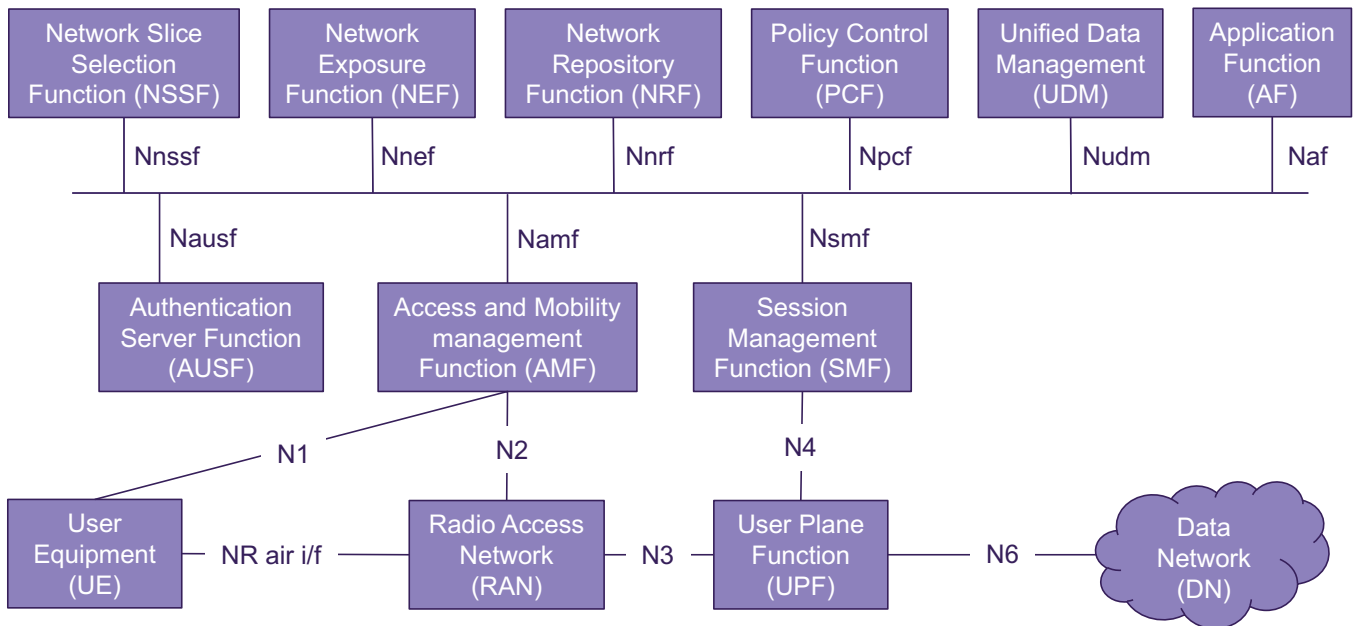


Figure 3: 5G service-based architecture

RAN and the core network although, as we'll discuss shortly, this isn't as simple in reality as the illustration in Figure 2 suggests. The N6 interface provides connectivity between the UPF and any internal or external networks or service platforms. This interface will include connectivity to the public Internet and will therefore contain the necessary Internet-facing firewalls and other smarts associated with the evolution of the Gi/SGi LAN¹ environment. The Gi/SGi LAN environment has evolved from GPRS through UMTS and LTE to provide a range of capabilities in support of mobile data network operation, including features such as Transmission Control Protocol optimisation, deep packet inspection and network address translation.

In addition to the familiar logical network diagram with defined interfaces, 3GPP has introduced an alternative view of the 5G network architecture which is known as Service Based Architecture (SBA). SBA takes advantage of recent developments in Network Functions Virtualisation and Software Defined Networking to propose a network based on virtualised infrastructure. This architecture will leverage service-based interactions between control plane functions as necessary. The solution will sit

on common computer hardware and call upon resources as required to manage demand at any instance. The use of SBA does not mandate a centralised solution; distributed computing could be implemented if appropriate. The SBA is illustrated in Figure 3.

3GPP states a number of principles and concepts for SBA (not all are exclusive to SBA), including:

- Separate control plane functions from user plane functions allowing independent scalability, evolution and flexible deployment.
- Modularise the functions design to enable flexible and efficient network slicing.
- Wherever possible, define procedures (the interactions between network functions) as services therefore their re-use is possible.
- Enable each network function to interact with other network functions directly, if required.
- Minimise the dependencies between the access network and core network; this

will enable different access types such as fixed broadband and WiFi (planned for future releases of 5G).

- Support a unified authentication function.
- Support stateless network functions such that the compute resource is decoupled from the storage resource.
- Support concurrent access to local and centralised services; this will enable support for low-latency services along with access to local data networks. To facilitate this, user plane functions can be deployed much closer to the access network.
- Support roaming with both home network routed traffic and local break-out traffic in the visited network.

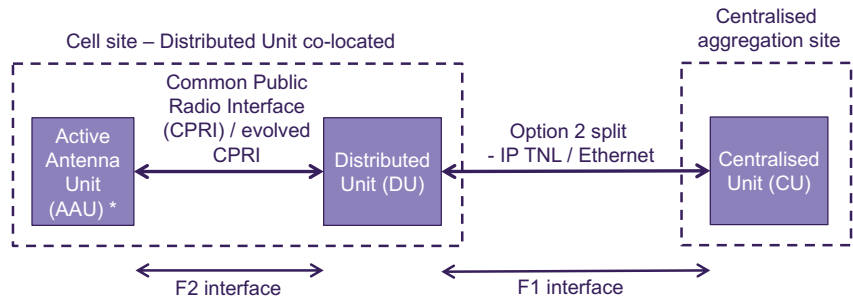
The SBA introduces a couple of functions that didn't exist in the traditional logical interface-based architecture representation; these are the Network Repository Function (NRF) and Network Exposure Function (NEF). The NRF provides control plane network functions with a mechanism to register and discover functionality so that next generation control plane network functions can discover each

¹The Gi-LAN interface is a 3GPP reference point between the mobile packet core and the packet data network or internet. In LTE networks the interface is referred to as the SGi-LAN and connects the Packet Gateway in the mobile core network to the packet data network.

other and communicate directly without making messages pass through a message interconnect function. The NEF receives information from other network functions (based on exposed capabilities of other network functions). It may store the received information as structured data using a standardised interface to a data storage network function. The stored information can be re-exposed by the NEF to other network functions and used for other purposes such as analytics. A practical example of use of the NEF is to aid the establishment of an application server-initiated communication with a UE where no existing data connection exists.

Functional decomposition of the RAN

In the high-level network architecture illustrated in Figures 2 and 3, the RAN is represented as a single functional entity whereas in reality the realisation of a 5G RAN is not so straightforward. In GSM/GPRS and UMTS there was a network controller which provided an interface between the radio access network and the core network. This network controller hid a lot of signalling from the core, particularly in UMTS, and managed a range of complex RAN functions. In LTE there is no network controller, the RAN manages a range of mobility management and radio optimisation activities between evolved Node Bs via the X2 interface. 5G effectively introduces a centralised RAN node albeit not a network controller as such. The 5G radio base station, known as a



*AAU illustrated; actual implementation could be AAU or passive antenna with Remote Radio Unit (RRU)

Figure 5: 5G RAN functional blocks and interfaces (excluding NR air interface)

next Generation Node B (gNB) is split into two entities: a gNB-Distributed Unit (gNB-DU (often shortened to DU)) and a gNB-Centralised Unit (gNB-CU (often shortened to CU)). The protocol layer interface at which this split will occur has been the topic of much debate in 3GPP and throughout the wider industry.

3GPP used the RAN protocol model illustrated in Figure 4 (3GPP TR 38.801) to discuss the functional split which should be implemented in 5G. Note that this protocol model is based on LTE as this was all that was known at the time although this doesn't differ significantly from 5G NR. The same terms are used although there have been some minor movements of functional sub-entities. Additionally, a new protocol, known as Service Data Adaptation Protocol (SDAP), has been introduced to the NR user plane to handle flow-based Quality of Service (QoS) framework in RAN, such as mapping between QoS flow and a data

radio bearer, and QoS flow ID marking.

Reading Figure 4 from left to right, Radio Resource Control (RRC) resides in the control plane while the data is user plane. As discussed above, SDAP will be inserted between data and the Packet Data Convergence Protocol (PDCP) for a standards-compliant 5G NR view of the protocol stack. Functions of PDCP include; IP header compression and decompression along with ciphering and deciphering (encryption of the data over the radio interface). PDCP feeds down the stack to the Radio Link Control (RLC) layer. RLC functions include; Error correction with Automatic Repeat request (ARQ), concatenation and segmentation, in sequence delivery and protocol error handling. Moving down the stack from RLC to the Medium Access Control (MAC) layer we find the following functions; multiplexing and de-multiplexing, measurement reports to RRC layer, Hybrid

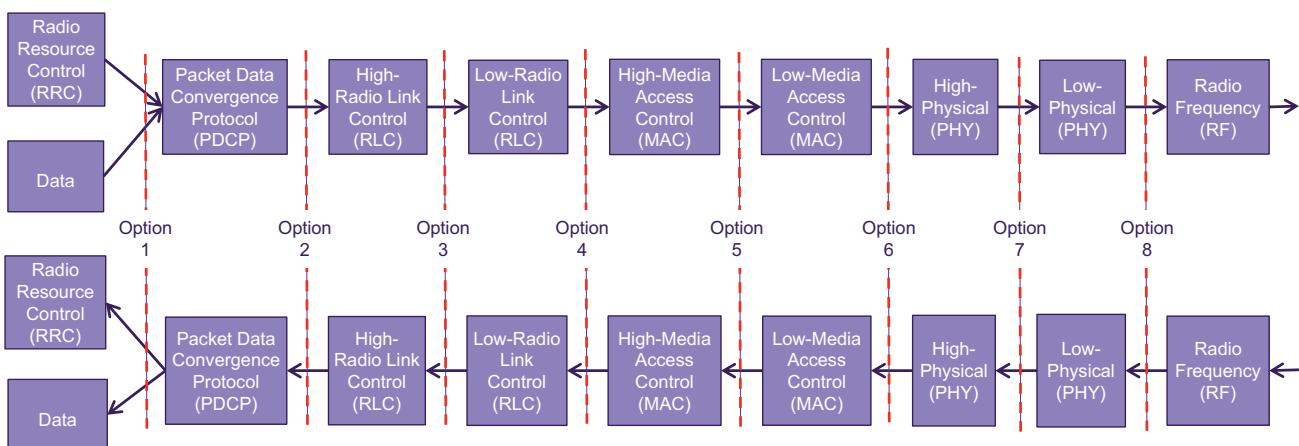


Figure 4: RAN protocol architecture as discussed in 3GPP TR 38.801

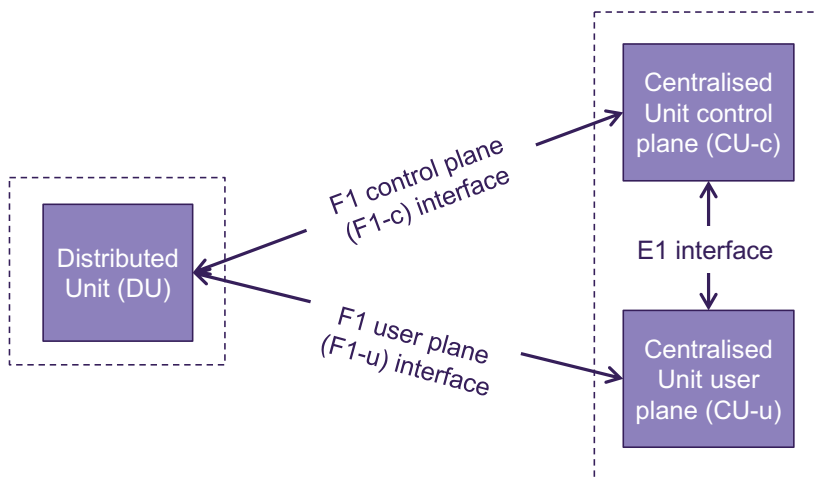


Figure 6: Control and user plane separation for F1 interface

ARQ error correction, scheduling and transport format selection. The physical layer takes care of the actual radio waveform and modulation scheme, amongst other things.

After much debate 3GPP agreed on an option 2 functional split, meaning that PDCP and therefore everything above this layer will reside in the CU while RLC and everything below it will reside in the DU. This is known as a higher layer split given its location within the protocol stack. The interface between the CU and DU has been designated “F1” (Figure 5); this will be supported by an IP transport network layer which will be carried over an underlying Carrier Ethernet network. Given the chosen location of the functional split there are no exacting latency requirements on the F1 interface; in fact it’s likely that the latency constraints applied to the F1 interface will be derived from the target service based latency. The data rate required for the option 2 F1 interface is very similar to that of traditional backhaul (LTE S1 interface as a reference) for a given amount of spectrum multiplied by the improved spectral efficiency of NR.

5G will build on the trend towards separate BaseBand Unit (BBU) and Remote Radio Unit (RRU) as increasingly deployed in today’s mobile networks. Additionally, 5G will introduce massive-MIMO antenna systems which will integrate the RRU functionality

within the antenna unit; these are known as Active Antennas Units (AAU). The interface between the current 4G BBU and RRU is based on the Common Public Radio Interface (CPRI) protocol in most implementations, CPRI is an option 8 interface as shown in Figure 4. Recent developments by the CPRI group, which consists of most of the major mobile RAN manufacturers, has resulted in an alternative lower layer split known as evolved CPRI (eCPRI). The initial implementation of eCPRI maps to an option 7 split. It is likely that both CPRI and eCPRI interfaces will be supported between DU and AAU as there are pros and cons to both approaches.

The interface between the DU and AAU is often referred to as the F2 interface (although this isn’t a formal 3GPP term, it may be adopted in the future). This may be local to the cell site or could be extended to form a more coordinated RAN. The challenge with extending this interface across a wide-area is the exacting performance requirements in terms of ultra-low latency and extremely high data transmission rates, particularly in the case of CPRI; eCPRI does benefit from some compression to reduce the data rate.

The theme of decoupling control and user plane is central to 5G network architecture development and therefore it is natural that this should be considered for the F1 interface.

The F1 interface is split into control and user plane interfaces which are known respectively as F1-c and F1-u as shown in Figure 6. The CU itself is also split into two functional entities; these could exist on the same hardware, on separate hardware on the same site or on separate hardware across different physical site locations. To connect the decomposed CU a new interface, designed E1, has been defined within 3GPP. The E1 interface (not to be confused with the legacy E1 (2048kbit/s) transmission interface) connects the CU-c and CU-u functional entities. The CU in its entirety can be built using virtualised infrastructure, as can the vast majority of the 5G network, the noticeable exception being certain radio frequency functions.

AUTHOR’S CONCLUSIONS

The strategic 5G network architecture comprises 5G NR and NGC although the latter is not likely to be deployed in the early years. In the first instance 5G will be supported alongside 4G on an EPC+ which is increasingly likely to be built on virtualised hardware and therefore a vEPC+. A NGC is necessary to realise the full feature set of 5G including the important concept of network slicing. The probable early deployment of 5G in the 3.5GHz band will require support from lower frequency bands to extend the range of the uplink to match the achievable downlink given effective Isotropic Radiated Power gains from beam-forming of signals enabled through the use of massive-MIMO antenna systems.

The initial enhanced uplink support is likely to come from LTE via a feature known as dual-connectivity although there are other NR-oriented proposals being studied within 3GPP, including 5G NR Carrier Aggregation and Supplemental Uplink. The functional decomposition of the RAN is an important aspect of the 5G network architecture; the location of DU and CU along with any potential split of CU-c and CU-u functions will require careful consideration to ensure an optimised network performance. It is

possible to co-locate all RAN functions and create a traditional fully distributed 5G base station (gNB) if particular use cases or deployment scenarios require this.

The increasing demands for ever higher peak and average data rates, greater area capacity density, lower-latency and enhanced performance will drive a more distributed next-generation core network. As functions of the RAN moves towards the core, certain core functions will move towards the RAN to facilitate services which are enabled from on-net infrastructure such as distributed user plane functions, Multi-Access Edge Computing and content distribution networks.

Chipset vendors are indicating that 5G NR-capable smartphones will be available from some manufacturers during mid-to-late 2019 and therefore it's likely that mainstream mobile-centric 5G network services will commence in many markets in and around the year 2020.

ABOUT THE AUTHOR

Andy Sutton is a Principal Network Architect within BT where he is responsible for 5G network architecture. He has



over 30 years of experience within the industry and is also engaged in the history and heritage of telecommunications. Andy holds an MSc in mobile communications and is a Visiting Professor with the School of Computing, Science and Engineering at the University of Salford, he is also a research mentor to the 5G Innovation Centre at the University of Surrey. Andy is a Chartered Engineer, Fellow of the IET, Fellow of the ITP and is a member of the Editorial Board for the ITP *Journal*.

Acknowledgments

The author would like to thank Maria Cuevas, Kevin Holley, Iain Stanbridge and John Whittington, all from BT TSO, for their valuable input to this paper.

ABBREVIATIONS

3GPP	3rd Generation Partnership Project	mMTC	Massive Machine Type Communications
AAU	Active Antenna Unit	NEF	Network Exposure Function
AF	Application Function	NGC	Next Generation Core
AMF	Access and Mobility management Function	NR	New Radio
ARQ	Automatic Repeat request	NRF	Network Repository Function
AUSF	Authentication Server Function	NSA	Non-Standalone
BBU	BaseBand Unit	NSSF	Network Slice Selection Function
CP-OFDM	Cyclic Prefix-Orthogonal Frequency Division Multiplex	PCF	Policy Control Function
CPRI	Evolved CPRI	PDCP	Packet Data Convergence Protocol
CU	Centralised Unit	QoS	Quality of Service
DN	Data Network	RAN	Radio Access Network
DU	Distributed Unit	RLC	Radio Link Control
eCPRI	Evolved CPRI	RRC	Radio Resource Control
eMBB	Enhanced Mobile Broadband	RRU	Remote Radio Unit
EPC	Evolved Packet Core	SBA	Service Based Architecture
FDD	Frequency Division Duplex	SDAP	Service Data Adaptation Protocol
gNB	next Generation Node B	SMF	Session Management Function
gNB-CU	gNB-Centralised Unit	TDD	Time Division Duplex
gNB-DU	gNB-Distributed Unit	TNL	Transport Network Layer
GPRS	General Packet Radio Service	UDM	Unified Data Management
GSM	Global System for Mobile Communications	UE	User Equipment
LTE	Long Term Evolution	UPF	User Plane Function
MAC	Medium Access Control	URLLC	Ultra-Reliable and Low Latency Communications
MIMO	Multiple Input, Multiple Output		

Join us at BT Centre on 16 May to hear more from Andy and other esteemed speakers on 5G. See page 7 of *Telecoms Professional* for more information.

ITP AUTHORS
 Want to know more? To contact the authors email your name, company name and email address to thejournal@theitp.org