Bartelt et al. EURASIP Journal on Wireless Communications and Networking (2017) 2017:89
DOI 10.1186/s13638-017-0874-7

EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

5G transport network requirements for the next generation fronthaul interface

J. Bartelt[1*], N. Vucic[2], D. Camps-Mur[3], E. Garcia-Villegas[4], I. Demirkol[4], A. Fehske[5], M. Grieger[5], A. Tzanakaki[6], J. Gutiérrez[7], E. Grass[7,8], G. Lyberopoulos[9] and G. Fettweis[1]

Abstract

To meet the requirements of 5G mobile networks, several radio access technologies, such as millimeter wave communications and massive MIMO, are being proposed. In addition, cloud radio access network (C-RAN) architectures are considered instrumental to fully exploit the capabilities of future 5G RANs. However, RAN centralization imposes stringent requirements on the transport network, which today are addressed with purpose-specific and expensive fronthaul links. As the demands on future access networks rise, so will the challenges in the fronthaul and backhaul segments. It is hence of fundamental importance to consider the design of transport networks alongside the definition of future access technologies to avoid the transport becoming a bottleneck. Therefore, we analyze in this work the impact that future RAN technologies will have on the transport network and on the design of the next generation fronthaul interface. To understand the especially important impact of varying user traffic, we utilize measurements from a real-world 4G network and, taking target 5G performance figures into account, extrapolate its statistics to a 5G scenario. With this, we derive both per-cell and aggregated data rate requirements for 5G transport networks. In addition, we show that the effect of statistical multiplexing is an important factor to reduce transport network capacity requirements and costs. Based on our investigations, we provide guidelines for the development of the 5G transport network architecture.

Keywords: 5G, Radio access network, New radio, Air interface, Fronthaul, Backhaul, Transport network, NGFI, Statistical multiplexing

1 Introduction

Cloud radio access network (RAN) (C-RAN) is considered to be one of the key technologies to increase efficiency and bring down costs in future mobile networks [1]. In C-RAN, baseband signal processing is offloaded from individual base stations—called remote units (RUs)—to a central unit (CU). This yields many benefits, such as simplified network maintenance, smaller form-factor RUs, efficient use of processing resources through statistical multiplexing at the CU, reduced costs for equipment rooms at base station sites, and spectral efficiency gains from joint processing such as coordinated multi-point (CoMP). On the other hand, it necessitates the deployment of a very demanding fronthaul (FH) network transporting the raw in-phase/quadrature-phase (I/Q) samples from the RUs to the CUs for processing. Commonly, this FH network is implemented based on the Common Public Radio Interface (CPRI) standard [2], which requires data rates of up to 24 Gbps per cell, a round-trip FH latency below 200 μs, low jitter, tight synchronization, and high reliability. These requirements can only be realized with high-capacity fiber or point-to-point wireless links, making the deployment of the FH network very costly, reducing the gains expected from centralization. The most promising approach to reduce the traffic load of the FH interface is the re-evaluation of the so-called functional split [3, 4], i.e., investigating which part of the signal processing can be performed at RUs and which part at CUs. By adopting only partial centralization, the FH requirements can be significantly relaxed while the main centralization benefits persist. These functional splits blur the difference between classical FH and backhaul (BH) networks, calling for converged transport networks that unify BH and FH equipment, and hence reduce deployment as well as operational costs. These networks can be facilitated by the

* Correspondence: jens.bartelt@tu-dresden.de
[1]Technische Universität Dresden, Vodafone Chair MNS, 01062 Dresden, Germany
Full list of author information is available at the end of the article

introduction of a next generation fronthaul interface (NGFI) [5].

C-RAN requirements and functional splits have been well investigated for 4G networks [3, 4] and are already considered for standardization by the IEEE 1914 working group [6] and in the CPRI Consortium's eCPRI standard [2]. However, for 5G networks, the introduction of new radio access technologies (RATs) is currently considered by 3GPP under the term new radio (NR) [7]. These new RATs, which could include technologies such as massive MIMO [8], millimeter wave (mmWave) communication [9], and non-orthogonal waveforms [10], will have a significant impact on the transport network. At the same time, new services, such as the Internet of things (IoT) [11], the Tactile Internet [12], or vehicular communication [13], will add new requirements such as ultra-low latency and extremely high availability. Hence, it is of fundamental importance to design the 5G transport network and the corresponding NGFI in view of the requirements these technologies induce.

Therefore, this article explores key 5G RATs and how their introduction will be reflected in the transport network. For this, we first revisit different functional splits currently under discussion for the NGFI and derive the corresponding peak transport network requirements. We then analyze the most promising 5G RATs to understand their impact on these requirements. For this, we utilize three exemplary configurations of 5G RATs that are aligned with the ongoing standardization as much as possible. To understand the behavior of real network traffic, we utilize live network measurements from a 4G network and extrapolate them to a 5G scenario. From this, we can derive statistical multiplexing gains, which will play a major role in reducing the required transport capacity. Based on these results, we give guidelines on how to design 5G transport networks.

2 5G radio access technologies and their impact on the transport network

2.1 Functional splits

The allocation of functions between the RU and CU, i.e., the functional split, has a major impact on the transport network and the corresponding NGFI requirements regarding data rate, latency, and synchronization. Figure 1 depicts a generic RAN signal processing chain of a mobile network. In principle, the split between RU and CU may be between any of the blocks depicted. In this work, we focus on three functional splits, which are capturing the most relevant trade-offs, denoted as A, B, and C[1] in Fig. 1. In general, splits higher up in the processing chain offer less centralization gains in terms of RRU size and cooperative processing, while reducing the requirements in terms of fronthaul data rate, latency, and synchronization. In the following, we summarize the most important features of these splits that will be relevant for the rest of the paper. As will be detailed, split A is very similar to the one currently employed in CPRI-based systems, offering full centralization at the price of tight requirements in terms of data rate and latency. Split B offers the benefit of featuring a FH data rate that depends on the load of the air interface, while split C is close to traditional BH, while still offering centralized medium access control (MAC)-layer functionalities like joint scheduling. For more details on the splits and their respective trade-offs, we refer the interested readers to [3, 4].

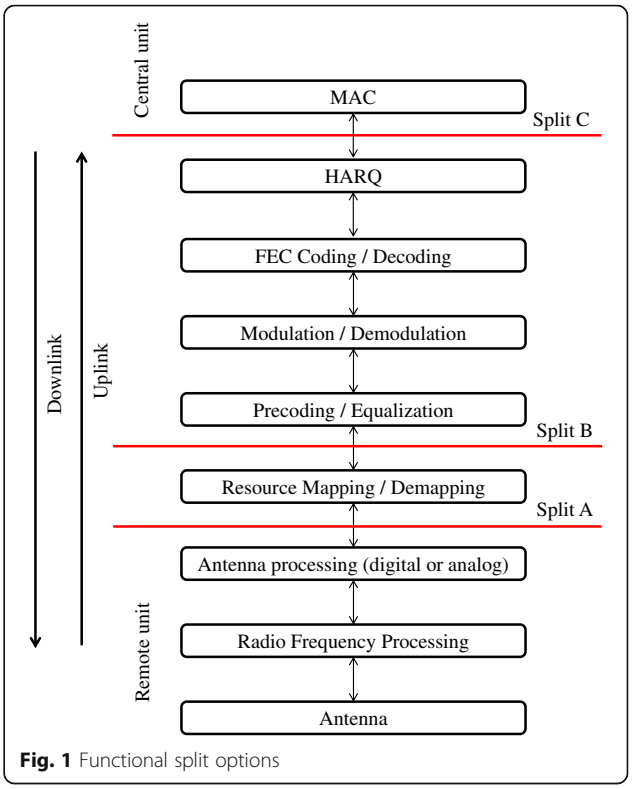

Fig. 1 Functional split options

2.1.1 Split A

Split A resembles the conventional C-RAN setup with a FH interface similar to CPRI. The only difference to CPRI is that we assume that some antenna processing is already performed at the RU, the reason for which will be explained in Section 2.2.3. The data forwarded between RUs and CUs consists of time-domain I/Q samples of constant rate which for this split can be calculated as

$$D_{\mathrm{A}} = 2 \cdot N_{\mathrm{L}} \cdot f_{\mathrm{S}} \cdot N_{\mathrm{Q,T}} \cdot \gamma, \tag{1}$$

with N_{L} being the number of analog-to-digital converter (ADC) chains, f_{S} the sampling frequency, $N_{Q,T}$ the

resolution of the time domain quantizer, and γ the transport overhead.

For this split, strict latency requirements of around 200 μs are induced by long-term evolution (LTE)'s hybrid automatic repeat request (HARQ) process [1]. However, this is a direct result of the LTE standard. A more fundamental limit to be considered is the channel's coherence time, as centralized functionalities such as adaptive modulation and coding scheme (MCS) selection, precoding, and scheduling require up-to-date channel information in order to perform adequately. From [14], an approximation of the channel coherence time in seconds can be obtained from

$$T_{\mathrm{C}} \approx \sqrt{\frac{9}{16\pi \mathrm{v}}} \frac{c}{\cdot f_{\mathrm{C}}}, \tag{2}$$

with c the speed of light in meter per second, v the UE's speed in meter per second, and f_C the carrier frequency in hertz. Note especially that the channel coherence time depends on the carrier frequency, which will play an important role for mmWave communication as discussed later.

In addition to these latency constraints, CPRI defines a synchronization requirement in the form of a delay estimation accuracy of 65 ns. This corresponds approximately to the duration of two samples of LTE, i.e., the delay estimation accuracy, which corresponds to the maximum allowed jitter, can be approximated by

$$T_{\mathrm{J}} \approx \frac{2}{f_{\mathrm{S}}}. \tag{3}$$

This requirement is motivated by the frequency diversity introduced by the timing offset between multiple antennas [15].

2.1.2 Split B

In split B, the mapping of data symbols to spectral resources is done at the RU. Since data symbols are only exchanged when data is available, the transport capacity is reduced and scales with the instantaneous cell load μ. The data rate for split B can accordingly be calculated as

$$D_{\mathrm{B}} = 2 \cdot N_{\mathrm{L}} \cdot N_{\mathrm{SC,act}} \cdot N_{\mathrm{Sy}} \cdot N_{\mathrm{Q,F}} \cdot T_{\mathrm{F}}^{-1} \cdot \mu \cdot \gamma, \tag{4}$$

with $N_{\mathrm{SC,act}}$ being the number of active subcarriers, N_{Sy} the number of symbols per frame, $N_{\mathrm{Q,F}}$ the resolution of the frequency domain quantizer, T_F the frame duration, and μ the utilization of the subcarriers, i.e., the load. The dependency on the actual load is a strong benefit of this split, as it enables statistical multiplexing, which will be further discussed in Section 3.3.

While split B still needs to meet the same latency constraints as split A, the synchronization on the transport network can potentially be relaxed as the different streams can be aligned at the RU. However, the synchronization between the antenna elements still needs to be ensured.

2.1.3 Split C

In split C, only higher MAC-layer functionalities (e.g., scheduling) are centralized, hence removing the option for joint PHY-layer processing such as joint transmission and reception in CoMP. The decentralization of the HARQ process, however, relaxes the latency requirements, which now depend only on channel coherence times. The transport data rate approximately follows the data rate experienced by the users, i.e., it depends on the used MCSs. In this regard, this functional split is very similar to BH links in current networks. The corresponding data rate can be calculated as

$$D_{\mathrm{C}} = N_{\mathrm{L}} \cdot N_{\mathrm{SC,act}} \cdot N_{\mathrm{Sy}} \cdot R_{\mathrm{c,MCS}} \cdot \log_2 M_{\mathrm{MCS}} \cdot T_{\mathrm{F}}^{-1} \cdot \mu \cdot \gamma, \tag{5}$$

with $R_{\mathrm{c,MCS}}$ being the code rate and M_{MCS} the number modulation symbols. Some of these splits are better suited for certain scenarios than others, e.g., depending on the availability of high capacity fiber or the density of access points. Hence, we foresee that multiple splits will coexist within the same 5G network. Accordingly, a converged network needs to be designed to handle the different traffic types induced by the functional splits, and the different RATs considered for 5G, which are being discussed next.

2.2 5G radio access technologies and use cases

To illustrate the impact of 5G RATs on future transport networks, we utilize three exemplary system configurations. The relevant downlink parameters as well as the resulting requirements in terms of data rate, latency, and synchronization are listed in Table 1 in comparison with a typical LTE system and will be further discussed in the following. Note that the specific 5G parameters are currently under standardization and might change with respect to this work. However, their general impact will made clear to be able to draw important conclusions on the overall network design.

2.2.1 Higher carrier frequencies

The adoption of higher carrier frequencies above 6 GHz, including the mmWave bands, is currently widely discussed as a possible enabler for 5G. 3GPP currently considers three bands for evaluation [7]: sub-6 GHz (around 2 GHz or around 4 GHz), around 30 GHz, and around 70 GHz, and, accordingly, we provide exemplary parametrizations for these three bands in Table 1. While higher carrier frequencies are foremost associated with higher bandwidths (see next paragraph), they also lead to lower

Table 1 Potential parametrization of a 5G radio access network using different frequency bands

Parameter	Symbol	Unit	4G	5G		
			LTE	Sub-6	Low mmWave	High mmWave
Carrier frequency	f_C	GHz	2	2	30	70
Channel size	BW	MHz	20	100	250	500
Sampling rate	f_S	MHz	30.72	150	375	750
# antennas	N_A	–	4	96	128	256
# ADC/DAC chains/layers	N_L	–	4	16	12	10
Overhead (control, line coding)	γ	–	1.33	1.33	1.33	1.33
Quantizer resolution time domain	$N_{Q,T}$	bit	15	15	12	10
Quantizer resolution frequency domain	$N_{Q,F}$	bit	9	9	8	7
Modulation order	M_{MCS}	–	64	1024	256	64
Max. code rate	$R_{C,MCS}$	–	0.85	0.85	0.85	0.85
Frame duration	T_F	ms	1	1	1	1
FFT size	N_{FFT}	–	2048	2048	2048	2048
# active subcarriers	$N_{SC,act}$	–	1200	1300	1300	1300
# data symbols per frame	N_{Sy}	–	14	70	150	300
Peak utilization	μ	–	1	1	1	1
Resulting requirements						
Channel coherence time at $v = 3$ km/h (2)	$T_{C,3}$	ms	76.17	76.17	5.08	2.18
Channel coherence time at $v = 50$ km/h (2)	$T_{C,50}$	ms	4.57	4.57	0.30	0.13
Channel coherence time at $v = 500$ km/h (2)	$T_{C,500}$	ms	0.46	0.46	0.03	0.01
Delay accuracy (3)	T_J	ns	65	13.33	5.33	2.67
Peak data rate split A (1)	D_A	Gbps	4.9	95.8	143.6	199.5
Peak data rate split B (4)	D_B	Gbps	1.6	34.9	49.8	72.6
Peak data rate split C (5)	D_C	Gbps	0.46	16.5	21.2	26.5

channel coherence times according to (2). As can be seen from Table 1, this will limit the tolerable FH delay to as little as 30 μs, as centralized precoding and adaptive coding and modulation needs up-to-date channel information in order to function properly. Hence, mmWave carriers will either require a decentralization of these functions, or even lower FH latencies need to be met compared with the ones currently specified in CPRI.

2.2.2 Larger bandwidths

Larger bandwidths are a major factor to enable higher data rates in 5G. This clearly increases the required transport data rates as well, regardless of the split used. The bandwidths considered in this paper (Table 1) are based on the 3GPP considerations [7], assuming that half of the available bandwidth is used for the downlink.

At the same time, the higher bandwidth constrains the delay accuracy in BH/FH due to significantly reduced sample duration, which is directly proportional to the bandwidth. Currently, CPRI has a two-way jitter requirement of 16.276 ns, which is about a quarter of the overall 65-ns time alignment requirement of 2G/3G/4G systems as given in Table 1 and Eq. (3). Looking at the requirements for the higher-bandwidth carriers, it can be observed that timing accuracies down to a few nanoseconds will have to be considered in 5G for centralized baseband processing.

2.2.3 Large antenna arrays

Utilization of large antenna arrays, both in sub-6 GHz and mmWave bands, is expected to be a key component of the envisaged 5G air interface [16]. Massive multi-user MIMO systems benefit from asymptotically orthogonal channels of different users, achieving significantly higher spectral efficiencies. In addition, mmWave systems require very large antenna arrays and corresponding beamforming techniques to overcome unfavorable wireless propagation conditions.

The direct application of CPRI-like FH consisting of I/Q samples for every antenna element (possibly several 100 s) leads to unacceptably high FH data rates. Therefore, only the data corresponding to independent spatial streams should be forwarded, as their number is typically much

lower than the number of antennas, thus reducing the transport capacity. As a consequence, data streams have to be mapped to the individual antenna elements at the RUs. While for sub-6 GHz systems, this mapping can be performed in the digital domain using precoders, for mmWave, the performance and power consumption of ADC/DAC chains encourages a partially analog approach, using a so-called hybrid beamforming architecture [17].

In Table 1, a maximum of 16 layers are considered for the 2-GHz carrier, while less layers are assumed for the mmWave carriers.

2.2.4 Higher order modulation

As of release 12 of LTE, modulation schemes up to 256-quadrature amplitude modulation (QAM) are supported on the air interface, and the WiFi standard 802.11ax already considers 1024-QAM. Increasing the modulation order is a straightforward method to increase the spectral efficiency and thus the air interface's capacity. As these modulation schemes provide an increased peak user data rate, they would also increase the transport data rate required for split C, since it linearly depends on the actual spectral efficiency of the air interface. The transport rates of splits A and B, in contrast, do not directly depend on the spectral efficiency. However, the higher order modulation schemes could ultimately require a higher quantization resolution, both in time and frequency domain, thus increasing the peak transport data rates of split A and B as well. The ADC resolution in Table 1 for the sub-6 GHz system is based on that of LTE systems. Lower resolutions can be expected for mmWave carriers due to the higher sampling rate and the associated power consumption. For frequency-domain samples, as present in split B, it has been observed that a lower resolution can be used [3].

2.2.5 New waveforms and frame structures

Although orthogonal frequency division multiplexing (OFDM) is still considered to be a strong contender for 5G networks, alternative waveforms have been investigated recently. Several multicarrier proposals employing filtering or pulse shaping exhibit better frequency localization properties than the standard OFDM and could be more suitable for use cases requiring flexible spectrum usage or asynchronous access [10]. In LTE systems, roughly 10% of the bandwidth is reserved for reducing adjacent channel interference. This percentage can be significantly reduced by utilizing the new waveforms, increasing the number of active subcarriers (e.g., in Table 1 from 1200 to 1300) and, correspondingly, the transport rates in higher functional splits such as B and C.

These new waveforms are also designed to be more flexible regarding frame and resource grid structure. Accordingly, the transport network will have to cope with varying data rates, which is currently not possible with the static data rate CPRI links.

Finally, the flexible frame structures will also lead to new protocol timings. As previously mentioned, especially the maximum HARQ delay defined for LTE currently limits the tolerable FH delay. If even stricter HARQ timings are introduced for 5G, this could severely limit the applicability of C-RAN architecture for 5G. This again highlights the importance of analyzing the impact of 5G RATs on the transport network already during RAT specification.

Single carrier (SC) waveforms might also play a prominent role in 5G, in particular for machine-type communication (MTC), uplink transmission, and mmWave communications due to better peak-to-average power ratio. In this study, we consider multicarrier modulation for all frequency ranges in order to provide a coherent framework. However, our main conclusions regarding the rate and latency hold for the SC waveforms as well.

2.2.6 Low-latency RAN and transport

A low end-to-end latency is a crucial requirement for the "Tactile Internet" [12]. Latencies as low as 1 ms, including RAN and transport network delays as well as application processing at the server and the user equipment (UE), shall enable a new generation of applications ranging from virtual reality to remote control to factory automation.

Regardless of the functional split, it is clear that the transport network's share of this latency should be kept to a minimum. While dedicated fiber, as used for CPRI deployments, achieves latencies of a few hundred microseconds (limited by propagation time), in packet-based networks, switching times and queueing significantly impact delay. In addition, the introduction of software defined networking (SDN) [18] in transport equipment (see also Section 4) poses new challenges in this regard. For example, traditional SDN architectures using reactive flow provisioning relay the first packet of each flow to the controller, which might incur unacceptable delays even for low demanding functional splits. Hence, transport SDN architectures based on proactive flow provisioning are required that can still provide the desired reliability and programmability features.

2.2.7 Machine-type communication

Massive MTC will be a major use case in 5G networks, enabling the IoT applications which encompasses a plurality of devices, such as sensors, smart meters, and cars. Although the traffic generated by many IoT applications will be comparatively low, it is very bursty, since few small data packets will be generated irregularly from potentially thousands of devices. Consequently, the total

traffic demand on the transport network is expected to be insignificant as compared to, e.g., high-definition video streaming. Thus, utilization-agnostic splits such as split A are highly inefficient for such applications due to their constant transport data rate. In addition, a fixed frame format could be seen as a drawback in a packet-based network, as small packets introduce a large overhead. Instead, a transport network supporting flexible frame formats could better deal with the nature of IoT traffic by, e.g., supporting different levels of packet aggregation.

For other applications such as vehicular communications, very high reliability will be of paramount importance. However, dedicated fiber links as currently used in CPRI-based deployment provide a single point of failure. Instead, a transport network is required in 5G which is able to provide alternative routes in case of outage. For example, ring topologies, which are challenging to implement with dedicated fiber, are quite common in, e.g., Ethernet networks, providing an effective 1 + 1 outage protection.

2.2.8 Network slicing

5G network slicing [16] is a novel architectural concept to enable the diverse set of use cases currently envisioned for 5G. Virtual end-to-end-networks or "slices" are created for each 5G service, which may require a different air interface per slice tailored to a given service, e.g., an "IoT slice" or a "tactile slice." Each 5G slice could thus benefit from a different functional split according to the service it provides. For example, in [19], an architecture for 5G RANs is proposed where the non-real-time components of the air interface are broken up into functions that can be virtualized and instantiated on a per-slice basis, following the paradigm of network function virtualization (NFV) [20]. Hence, 5G slicing will build on NFV to instantiate per-slice virtual functions, as well as on SDN, to connect the different functions belonging to a given slice. In [21], a cellular architecture combining NFV and SDN is proposed that enables flexible service definition over LTE networks. Similar architectural concepts are likely to be applicable to 5G to define end-to-end network slices. Finally, it is worth highlighting that the impact of adopting SDN/NFV on the security of 5G transport networks deserves special attention and is an area of ongoing research [22].

3 5G transport requirements

3.1 Peak requirements

The peak requirements per cell of the 5G technologies discussed above are summarized at the bottom of Table 1 and are illustrated in Fig. 2. Note that while the parameters of 5G RATs are currently not standardized, the numbers provided are a realistic outlook to the transport requirements if the RATs discussed are indeed implemented. In addition, not all of the technologies might be implemented in the first release of 5G but rather at later stages. Note for example that the data rate given for split C is between 16 and 26 Gbps. Considering that split C is close to the data rates experienced by the user but still includes MAC and transport overhead, these data rates are in line with the capacity of up to 10 Gbps considered for 5G. Looking at the delay accuracy and channel coherence times, we can see that they differ by orders of magnitude depending on the different RATs. This already indicates that future transport networks will have to deal with a large variety of requirements, which will be further discussed in Section 4.

For the lower functional splits, the data rates range from 35 to almost 200 Gbps per cell. However, these are peak requirements. We have already remarked that a key benefit of splits B and C is that the associated transport data rates scale with the actual utilization of resources on the access link. Thus, when using these splits, dimensioning the transport network according to peak data rates is not efficient in most cases. Instead, the next generation of mobile networks (NGMN) alliance has derived guidelines for dimensioning transport networks based on busy hour utilization, which leverage the statistical multiplexing occurring in practical networks [23]. Hence, it is important to understand the transport requirements introduced by the 5G RATs and functional

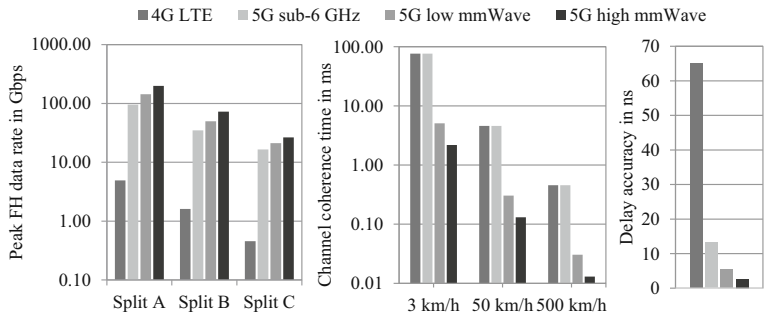

Fig. 2 Fronthaul requirements

splits considered under practical network conditions as done in the following.

3.2 Live network measurements

To this effect, we combine the 5G RAT configurations from Table 1 with real downlink measurements from a real-life LTE network. The measurements were collected from 10 LTE sites with $C = 33$ cells in total, covering an area of about 2.5 km^2 in downtown Athens, Greece. Twenty-seven of the 33 cells utilize 20 MHz bandwidth, while the other six cells utilize 10 MHz. The measurements were taken on a 15-min basis for a time period of 15 days. The measurements taken include (for each cell c and 15-min time instance $t = (d,m)$, d being the day and m the 15-min interval):

- The percentage of utilized physical resource blocks (PRBs) relative to the total number of PRBs, $\mu(c,t)$
- The MCS distribution, $p(MCS_{i,c,t})$, with $i = \ldots 28$ being the index of the 28 MCS utilized in LTE
- The maximum cell throughput, $D_{\max}(c,t)$
- The maximum number of UEs per cell, $N_{\mathrm{UE}}(c,t)$

From this, Fig. 3a shows the maximum and average utilization measured, i.e.,

$$\mu_{\max}(t) = \max_c \mu(c,t), \tag{6}$$

and

$$\mu_{\mathrm{avrg}}(t) = \frac{1}{C}\sum_{\mathrm{c}} \mu(c,t), \tag{7}$$

as well as the MCS distribution $p(MCS_i,c,t)$ of the MCS 0 to 28.

While these measurements are based on a 4G network, we assume that similar traffic patterns will be observed for the mobile broadband use case of 5G, as general user behavior should stay the same.[2] Furthermore, it can be seen that the average utilization in the measurements is quite low. In order to project a 5G traffic distribution, the load is scaled as described in the following. It is assumed that each user has an average traffic demand per user of $D_{5G} = 300$ Mbps as recommended for the downlink by the NGMN alliance for broadband access in dense areas in [16] and that the spectral efficiency of a 5G network would be approximately five times higher than in a current LTE network. Accordingly, the loads for a high load scenario are defined as

$$\mu'(c,t) = \mu(c,t)\frac{D_{5\mathrm{G}}}{D_{\max}(c,t)/N_{\mathrm{UE}}(c,t)\cdot 5}. \tag{8}$$

The resulting utilization is shown in Fig. 3b. Thus, we hereafter use the measured utilization μ ("low load") and scaled utilization μ' ("high load") to provide exemplary transport data rate requirements of 5G networks.

In order to dimension the transport network appropriately, we now consider the busy hour traffic following the recommendations introduced by the NGMN alliance in [23]. The busy hour was selected as the hour with the largest utilization sum over all cells, i.e.,

$$t_{\mathrm{busy}} = \underset{m}{\mathrm{argmax}} \sum_c \sum_d \sum_m^{m+3} \mu(c, t = (d,m)) \,. \tag{9}$$

The busy hour was found to be from 12:15 to 13:15 h. From this, a busy hour MCS distribution was calculated as the average MCS distribution in that hour, i.e.,

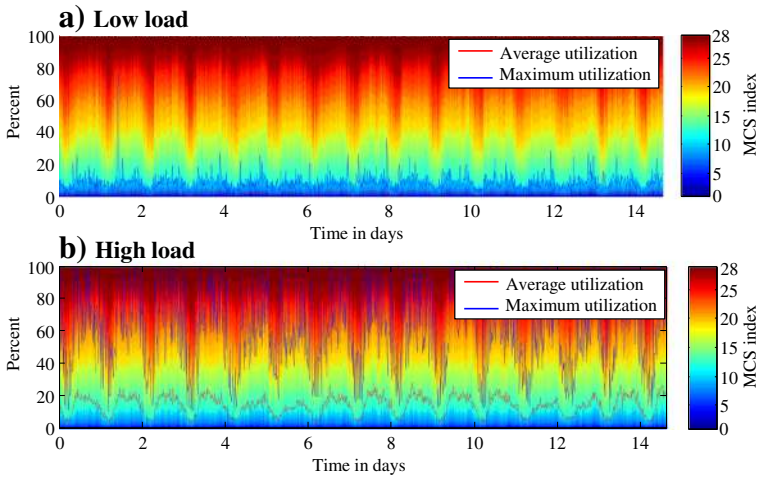

Fig. 3 Measured MCS distribution and maximum and average resource utilization for **a** low network loads and **b** high network loads

$$p_{\text{busy}}(MCS_i) = \frac{1}{C \cdot D} \sum_c \sum_d p(MCS_i, c, t_{\text{busy}}). \quad (10)$$

The resulting distribution is depicted in Fig. 4a. Note that MCS 10 and 17 are very close to their adjacent MCS in terms of spectral efficiency and are hence not utilized in this network. Accordingly, the figure shows a probability of zero for these two MCS.

All loads observed in the busy hour were used to accumulate a discrete busy hour load distribution as

$$p_{\text{busy}}\left(\mu \in \bigcup_{d,c} \mu(c, t_{\text{busy}})\right) = \frac{1}{4 \cdot C \cdot D}. \quad (11)$$

with the factor 4 accounting for the four 15-min traces per hour. The cumulative distribution function (CDF) of the resulting load distributions are illustrated in Fig. 4b.

Using the busy hour utilization and MCS distribution, it is possible to calculate the probability distribution of the data rates of split A, B, and C and for the three different RATs introduced in Table 1. For this, all combinations of occurring loads μ and MCS_i are inserted into Eqs. (1), (4), and (5) to obtain the data rates $D(\mu, MCS_i)$. Assuming for simplicity that there is no correlation between a certain load and an MCS, the probability of each data rate to occur can then be calculated as

$$p_{\text{busy}}\left(D(\mu, MCS_i)\right) = p_{\text{busy}}(\mu) \cdot p_{\text{busy}}(MCS_i). \quad (12)$$

To account for the fact that modulation schemes of up to 1024-QAM were assumed instead of the 64-QAM in the actual measurements, the spectral efficiency of each MCS was scaled accordingly by up to 10/6. The resulting complementary CDFs (CCDFs) of the different data rates are depicted in Fig. 5. Note that the data rates of the different RATs are normalized to their peak value (i.e., the values from Table 1) and are hence the same for all three RATs. While the rate is constant for split A, the rates of split B depend on utilization, and those for split C depend both on utilization and on the MCS distribution. We can also see that, e.g., in the low load scenario for split C, the data rate exceeds 30% of the peak rate for only 1% of the time. This fact gives rise to statistical multiplexing which will be discussed in the next section.

3.3 Statistical multiplexing

Statistical multiplexing can be exploited when two factors are considered: the aggregation of transport streams of several cells and the introduction of a certain outage rate.

First, an outage rate can be considered as the transport network is commonly not dimensioned for peak rates in order to save costs [24]. Instead, it is dimensioned for a certain percentile, e.g., the 95th percentile Q_{95}, i.e., the offered traffic can be transported without queueing with a probability of 95%:

$$p(D \leq Q_{95}) = 0.95. \quad (13)$$

As an example, consider split C of the low load scenario in Fig. 5: here, the Q_{95} is only 17% of the peak data rate Q_{100}.

Second, when the transport streams of several base stations are aggregated onto one link, the resulting data rate distribution is given as the C-fold convolution of the individual distributions $p(D)$. Accordingly, the percentiles $Q(C)$ of C cells are different from a simply scaled percentile of one cell $C\ Q(1)$. This combination of outage rate and aggregation of several RUs yields what we call the statistical multiplexing gain. Such gain is illustrated in Fig. 6 for an example of a uniform load distribution. The figure shows the resulting aggregated data rates for the 2-GHz carrier and split B when 1, 2, 4, and 8 cells are aggregated. Note how the uniform distribution becomes increasingly long-tailed; in other words, it is highly unlikely that several cells will exhibit peak rates at the same time instance. In the figure, we also highlight the aggregated peak rates ($C\ Q_{100}(1)$), aggregated 95th percentile without accounting for statistical

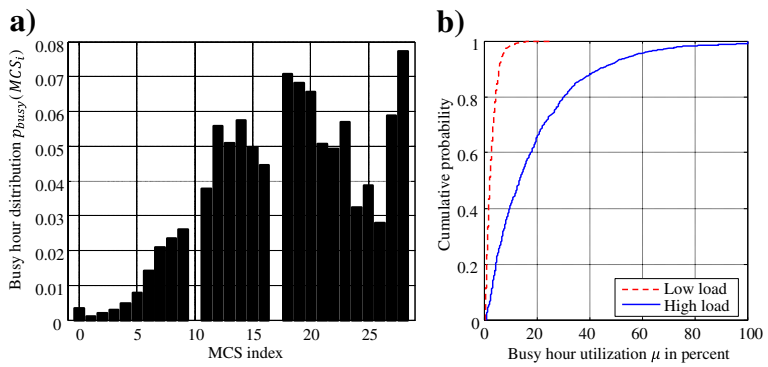

Fig. 4 Busy hour distributions. **a** MCS index distribution. **b** CDF of utilization

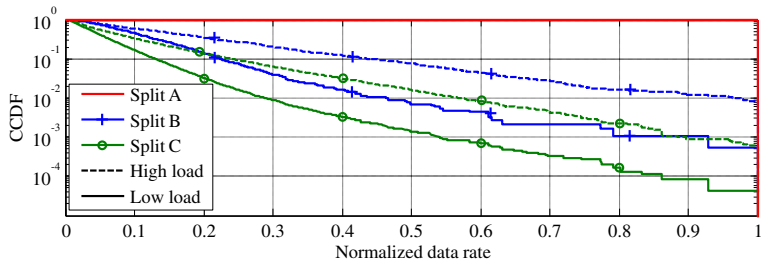

Fig. 5 CCDF of normalized data rates when employing split A, B, and C

multiplexing, C $Q_{95}(1)$, and the 95% percentile of all aggregated cells when considering statistical multiplexing ($Q_{95}(C)$) for $C = 8$ cells. These values illustrate how both the outage percentile and the aggregation of several cells can decrease the required transport data rate on an aggregation link dramatically. As the statistical multiplexing effect cannot be relied on for only one or a few cells, we follow the methodology from [23] and calculate the required capacity as the maximum between the 95th percentile of C cells and the peak capacity required for one cell. We hence define the required data rates and the statistical multiplexing gain as

$$R_{\text{req,nomux}}(C) = C \cdot Q_{100}(1) \quad \text{without multiplexing,} \tag{14}$$

$$R_{\text{req,mux}}(C) = \max(Q_{100}(1),\ Q_{95}(C)) \text{ with multiplexing,} \tag{15}$$

$$g_{\text{mux}}(C) = \frac{R_{\text{req,mux}}(C)}{R_{\text{req,nomux}}(C)} \quad \text{multiplexing gain.} \tag{16}$$

Figure 7 illustrates the resulting capacity to be deployed in aggregation networks for the different functional splits, for RATs, and for both the high load and low load scenarios. We give the capacity of up to 1000 cells, which we consider to be the approximate order of magnitude for the number of cells in a single large city.[3] In addition, the corresponding multiplexing gains for the high-load scenario are given in Fig. 8.

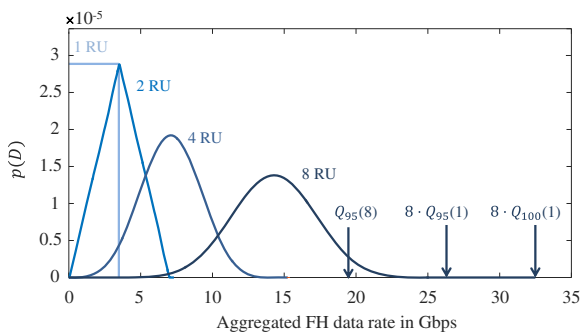

Fig. 6 Convolution of PDFs leading to multiplexing gains for split B and 2-GHz carrier

First, note that according to the overall increase in data rates considered for 5G networks, the transport data rates are also expected to increase by order of magnitude and can easily reach the Tbps range. However, the introduction of the lower functional splits B and C can mitigate this effect, and statistical multiplexing can further reduce the required capacity by almost ten times as seen in Fig. 8. The degree of statistical multiplexing gain depends on the number of aggregated cells but shows already high values for a few dozens of cells. In addition, split C, which exhibits the highest variability of per-cell traffic due to its dependence on both the overall load and the utilized MCS, shows the highest gains, while the static rates of split A do not offer any statistical multiplexing. These factors need to be considered in the design of future transport networks in order to make them economically feasible, which will be discussed in the next section.

4 Design guidelines for 5G transport networks

From the analysis and results in the previous sections, important conclusions can be drawn towards the design of 5G transport networks, which will be discussed next. This covers the necessity of converging fronthaul and backhaul networks, the technologies that can enable this convergence, and finally a change in network management towards SDN.

4.1 Transport network convergence

The 5G NGFI capacity requirements derived in Section 3 indicate that full centralization of baseband processing is not reasonable for all air interface technologies. The introduction of bandwidth in the GHz range, as well as large antenna arrays, increase transport data rates beyond what can be supported with current or near-future transport technology at feasible costs. Therefore, functional splits, where the transport rates are proportional to the actual utilization, should be considered for future 5G systems.

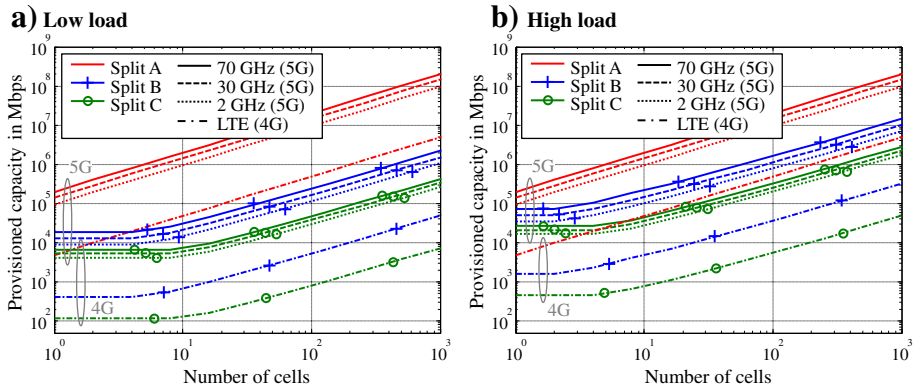

Fig. 7 Provisioned transport capacity versus number of cells for **a** low-load scenario and **b** high-load scenario

We also showed in the previous section how statistical multiplexing can also contribute to greatly reduce the required capacity on aggregation links. We would like to note that the statistical multiplexing gain will also greatly depend on the variance of the transported traffic in addition to the number of aggregated cells. The above analysis included only traces from an urban environment. It can be expected that aggregating more diverse types of cells will increase the variance of the traffic and—correspondingly—the multiplexing gain. Accordingly, future transport networks should aim to aggregate as diverse cells as possible, multiplexing, e.g., small and macro cells or urban and rural ones. Furthermore, recall that the used traces were performed on a 15-min time scale, which basically constitutes an averaging over 15 min. Hence, it can be expected that multiplexing on the timescale of a packet duration would further increase the traffic variance and, subsequently, the statistical multiplexing gain. In addition, the 5G transport network will have to support different traffic types along with varying requirements, not only according to the different splits and air interfaces but also depending on the use cases. For example, while massive MTC applications might require very low data rates and can deal with a certain degree of packet loss, a low-latency and resilient transport network will be required for other applications. Flexible waveforms with variable bandwidth, frame structure, and data rates should also be reflected in the transport network. Flexibility is hence a key aspect for future transport which could be achieved by utilizing packet switching. Since packet-based networks are currently in use for BH traffic, this calls for a convergence of FH and BH networks, transporting different types of traffic over a shared infrastructure, to reduce hardware diversity and to share resources, while simplifying management and deployment. However, currently used technologies will have to be adapted to support packet switching.

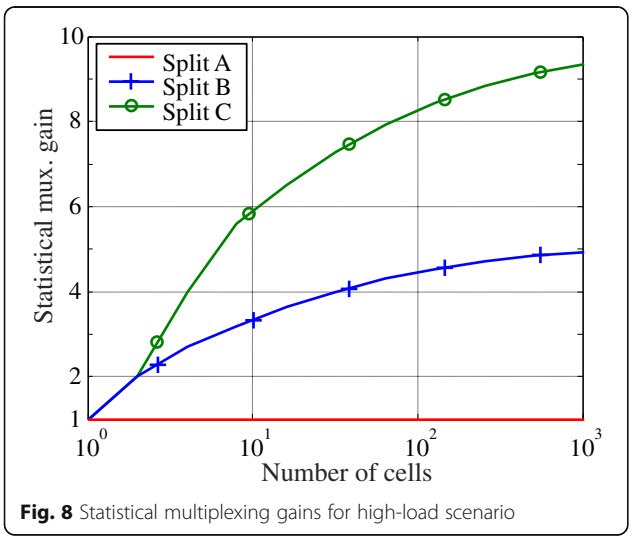

Fig. 8 Statistical multiplexing gains for high-load scenario

4.2 Enabling transport technologies

Dedicated fiber connectivity, currently favored for CPRI-based FH networks, does not offer resource sharing between different fibers or wavelengths. Packet-switched networks based on Ethernet are the most promising alternative as reflected by recent initiatives such as the IEEE 1914 Working Group [6]. Currently, synchronization is still challenging in packet-based networks and jitter introduced by switches and queues could become critical for FH traffic. In this regard, technologies such as Synchronous Ethernet [25] and the Precision Time Protocol (IEEE 1588) [26], which are currently investigated by, e.g., the IEEE 802.1 Time Sensitive Networking Task Group [27], are good candidates to efficiently support FH over Ethernet.

Supporting variable data rates is also a particular challenge at the physical level, due to the inflexible, dedicated links currently utilized. A packet-based flexible optical transport will comprise both passive and active solutions. Passive optical network (PON) solutions will

be based on wavelength division multiplexing (WDM)-PONs, while active solutions will adopt more flexible and dynamic WDM technologies such as the time-shared optical network [28], which enable very granular sub-wavelength bandwidth allocation that is a key to efficiently utilize optical bandwidth in the aggregation/metro segment of converged FH/BH networks. Wireless transport technologies are also considered as candidate technologies for future transport networks due to their lower cost and higher flexibility compared to fiber.

Recently, drone-based mobile networks have been proposed in order to provide additional coverage to support temporary high traffic demand or unexpected or critical scenarios [29]. While still in early stages of research, it can already be foreseen that such networks will pose new challenges on the transport network, with both wireless and wired solutions (via a tether, see [30]) having been proposed. However, such aerial networks could also provide new solutions for future transport networks, e.g., by acting as relays or providing additional transport capacity via aerial routes in times of unexpected high transport traffic or failure of ground-based transport equipment.

4.3 Quality of service and management

Current CPRI-based transport networks consider just two types of traffic to be transported: data and control signals. Synchronization is directly supplied via CPRI line rate. Given the variety of potential 5G RATs, applications, and functional splits, many different types of traffic will have to be transported in 5G transport networks, each with their own requirements. While split A traffic will require low latencies, split C traffic can typically cope with much relaxed latencies. On the other hand, Tactile Internet traffic will require low latencies even for split C. To transport these different types of traffic over a unified network with singular requirements based on the strictest use case would clearly be cost-inefficient. A future transport network will hence have to support streams with different qualities of service. Packet-based networking can here also help to facilitate this via packet prioritization. However, this will make the management of the network more challenging, calling for SDN solutions.

SDN [18] is a recent networking paradigm, which separates control and data planes to enhance flexibility and to achieve programmability of network technologies. SDN is a key enabler for converged FH/BH networks in 5G. It will be used to virtualize the transport network in order to support slicing and to allow a flexible deployment of virtual functions in different places of the network, as is required for the support of flexible functional splits. In some cases, a separate out-of-band network for SDN signaling traffic may be too expensive to maintain, or reliability constraints may require a certain degree of distributed control to be kept in the network elements, thus, balancing between distributed and centralized control. Therefore, further research is needed to holistically apply the SDN paradigm to transport networks.

5 Conclusions

5G mobile networks will utilize a diverse set of access technologies, which will increase FH and BH requirements dramatically. Furthermore, multiple RATs with different degrees of centralization as well as novel use cases must be supported at a reasonable cost. The impact of those developments must be taken into account for the design of 5G transport networks and the corresponding next generation fronthaul interface. We showed how new functional splits can avoid a costly increase in necessary transport capacity which, otherwise, would be required by new technologies like mmWave and massive MIMO. By analyzing traffic measurements from a real-life, commercial network, we were able to illustrate that the combination of utilization-dependent functional splits and statistical multiplexing can additionally reduce aggregated transport traffic by up to almost one order of magnitude as compared to today's CPRI-based networks. However, this will require a more flexible, dynamically configurable transport network, which can transport different types of traffic. To enable the required degree of flexibility while keeping the network manageable, a converged, packet-based, and SDN-enabled transport network will be required to support 5G radio access networks.

6 Endnotes

[1]Please note that the nomenclature of these splits is not harmonized across different works, e.g., split C of this work might be called split E in another work.

[2]Note that different traffic patterns might emerge, e.g., for machine-to-machine communication, see, e.g., [31]. A detailed study of 5G traffic patterns we leave for future work.

[3]One thousand single-cell base stations with an inter site distance of 500 m in a hexagonal layout approximately correspond to an area of 15 × 15 km.

Abbreviations

ADC: Analog-to-digital converter; BH: Backhaul; CCDF: Complementary cumulative distribution function; CDF: Cumulative distribution function; CoMP: Coordinated multi-point; CPRI: Common Public Radio Interface; C-RAN: Cloud radio access network; CU: Central unit; FH: Fronthaul; HARQ: Hybrid automatic repeat request; I/Q: In-phase/quadrature-phase; IEEE: Institute of electrical and electronics engineers; IoT: Internet of things; LTE: Long-term evolution; MAC: Medium access control; MCS: Modulation and coding scheme; MIMO: Multiple input multiple output; mmWave: Millimeter wave; MTC: Machine-type communication; NFV: Network function virtualization; NGFI: Next generation fronthaul interface; NGMN: Next generation mobile networks; OFDM: Orthogonal frequency division multiplexing; PDF: Probability density function; PON: Passive optical network; PRB: Physical resource block; QAM: Quadrature amplitude modulation; RAN: Radio access network; RAT: Radio access technology; RU: Remote unit; SC: Single carrier; SDN: Software defined network; UE: User equipment; WDM: Wavelength division multiplexing

Acknowledgements
The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 671551 (5G-XHaul). The European Union and its agencies are not liable or otherwise responsible for the contents of this document; its content reflects the view of its authors only. The authors would like to thank Eleni Theodoropoulou, Markos P. Anastasopoulos, and Dimitra Simeonidou for their contributions to this work.

Availability of data and materials
The dataset supporting the conclusions of this article cannot be made available, as it could disclose corporate secrets of one or several authors' institutions.

Authors' contributions
JB coordinated the writing of the overall document and contributed Section 3.3. NV contributed Sections 2.2.1, 2.2.2, and 2.2.5, as well as the parametrization in Table 1. D. CM contributed Sections 2.2.7 and 2.2.8, as well as to extrapolating the measurement data in Section 3 to the 5G case. EGV contributed to the evaluation of the measurement data in Section 3.2. ID contributed by extracting the individual statistics from the measurements in Section 3.2. AF contributed Section 2.1. MG contributed Sections 3.1 and 2.2.3. AT contributed Section 4.2. JG contributed Section 1. EG contributed Section 5 and the abstract. GL performed the measurements that are the basis for Section 3 and contributed to the interpretation of their results. GF contributed Section 2.2.6, as well as to the overall concept of the study, the necessary measurements, their interpretation, and corresponding conclusions. In addition to the contributions listed above, all authors provided extensive reviews to the overall manuscript and, being part of a larger project, contributed to the conception of the work and the interpretation of the results. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Technische Universität Dresden, Vodafone Chair MNS, 01062 Dresden, Germany. [2]Huawei Technologies Duesseldorf GmbH, Riesstrasse 25, 80992 Munich, Germany. [3]i2CAT Foundation, Gran Capità, 2-4, Nexus I Building, 08034 Barcelona, Spain. [4]Universitat Politècnica de Catalunya, c/Jordi Girona 1-3, Modul C3, 08034 Barcelona, Spain. [5]Airrays GmbH, Kramergasse 4, 01067 Dresden, Germany. [6]University of Bristol, Merchant Venturers' Building, Woodland Road, Clifton, Bristol BS8 1UB, UK. [7]IHP GmbH, Im Technologiepark 25, 15236 Frankfurt (Oder), Germany. [8]Institut für Informatik, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany. [9]COSMOTE Mobile Telecommunications S.A., 44, Kifissias Ave., 15124 Maroussi, Athens, Greece.

Received: 18 November 2016 Accepted: 2 May 2017
Published online: 15 May 2017

References

1. A Checko et al., Cloud RAN for mobile networks—a technology overview. IEEE Commun Surv Tutorials **17**(1), 405–426 (2015)
2. Common Public Radio Interface, http://www.cpri.info/. Accessed 1 Aug 2016
3. U Dötsch et al., Quantitative analysis of split base station processing and determination of advantageous architectures for LTE. Bell Labs Tech J **18**(1), 105–128 (2013)
4. D Wübben et al., Benefits and impact of cloud computing on 5G signal processing. IEEE Signal Process Mag **31**(6), 35–44 (2014)
5. I CL, Y Yuan, J Huang, S Ma, C Cui, R Duan, Rethink fronthaul for soft RAN. IEEE Commun Mag **53**(9), 82–88 (2015)
6. IEEE Next generation fronthaul interface (1914) working group, http://sites.ieee.org/sagroups-1914/. Accessed 1 Oct 2016.
7. 3GPP, TR 38.913, *Study on scenarios and requirements for next generation access technologies, 3GPP technical report*, 2016
8. EG Larsson, O Edfors, F Tufvesson, TL Marzetta, Massive MIMO for next generation wireless systems. IEEE Commun Mag **52**(2), 186–195 (2014)
9. TS Rappaport et al., Millimeter wave mobile communications for 5G cellular: it will work!, IEEE Access **1**, 335-349 (2013)
10. G Wunder et al., 5GNOW: non-orthogonal, asynchronous waveforms for future mobile applications. IEEE Commun Mag **52**(2), 97–105 (2014)
11. L Atzori, A Iera, G Morabito, The Internet of things: a survey. Comput Netw **54**(15), 2787–2805 (2010)
12. G Fettweis, The Tactile Internet: applications and challenges. IEEE Veh Technol Mag **9**(1), 64–70 (2014)
13. F Qu, FY Wang, L Yang, Intelligent transportation spaces: vehicles, traffic, communications, and beyond. IEEE Commun Mag **48**(11), 136–142 (2010)
14. TS Rappaport, Wireless communications: principles and practice, 2nd ed., (Prentice Hall, Upper Saddle River, NJ, 2002).
15. Ericsson, R1-072463: absence of array calibration—impact on precoding performance, written contribution to TSG-RAN WG #49 (2007), http://www.3gpp.org/DynaReport/TDocExMtg–R1-49–26034.htm. Accessed 09 Sep 2016.
16. NGMN Alliance, *NGMN 5G white paper, white paper*, 2015. http://ngmn.org/uploads/media/NGMN_5G_White_Paper_V1_0.pdf. Accessed 1 Aug 2016
17. TE Bogale, LB Le, *Beamforming for multiuser massive MIMO systems: digital versus hybrid analog-digital, IEEE Global Communications Conference, Austin, TX*, 2014
18. S Sezer et al., Are we ready for SDN? Implementation challenges for software-defined networks. IEEE Commun Mag **51**(7), 36–43 (2013)
19. P Marsch et al., 5G radio access network architecture: design guidelines and key considerations. IEEE Commun Mag **54**(11), 24–32 (2016)
20. B Han, V Gopalakrishnan, L Ji, S Lee, Network function virtualization: challenges and opportunities for innovations. IEEE Commun Mag **53**(2), 90–97 (2015)
21. A Bradai, K Singh, T Ahmed, T Rasheed, Cellular software defined networking: a framework. IEEE Commun Mag **53**(6), 36–43 (2015)
22. H Hawilo, A Shami, M Mirahmadi, R Asal, NFV: state of the art, challenges, and implementation in next generation mobile networks (vEPC). IEEE Netw **28**(6), 18–26 (2014)
23. NGMN Alliance, *Guidelines for LTE backhaul traffic estimation, white paper*, 2011. http://www.ngmn.de/uploads/media/NGMN_Whitepaper_Guideline_for_LTE_Backhaul_Traffic_Estimation.pdf. Accessed 1 Aug 2016
24. Y d'Halluin, PA Forsyth, KR Vetzal, Wireless network capacity management: a real options approach. Eur J Oper Res **176**(1), 584–609 (2007)
25. JL Ferrant et al., Synchronous Ethernet: a method to transport synchronization. IEEE Commun Mag **46**(9), 126–134 (2008)
26. IEEE Std, *1588-2008: standard for a precision clock synchronization protocol for networked measurement and control systems*, 2008
27. IEEE 802.1 Time Sensitive Networking Task Group, http://www.ieee802.org/1/pages/tsn.html. Accessed 1 Oct 2016
28. BR Rofoee et al., *First demonstration of service-differentiated communications over converged optical sub-wavelength and LTE/WiFi networks using GEANT link, Optical Fiber Communication Conference, Los Angeles, CA*, 2015
29. I Bor-Yaliniz, H Yanikomeroglu, The new frontier in RAN heterogeneity: multi-tier drone-cells. IEEE Commun Mag **54**(11), 48–55 (2016)
30. S Chandrasekharan et al., Designing and implementing future aerial communication networks. IEEE Commun Mag **54**(5), 26–34 (2016)
31. MZ Shafiq et al., A first look at cellular machine-to-machine traffic: large scale measurement and characterization. ACM SIGMETRICS Perform Eval Rev **40**(1), 65–76 (2012)