Peer-reviewed version available at *Information* **2018**, *9*, 72; doi:10.3390/info9040072

Article

Random Linear Network Coding for 5G Mobile Video Delivery

Dejan Vukobratovic[1], Andrea Tassi[2], Savo Delic[1] and Chadi Khirallah[3]

[1] University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia; dejanv@uns.ac.rs
[2] University of Bristol, Department of Electronic and Electrical Engineering, Bristol, UK; a.tassi@bristol.ac.uk
[3] University of Edinburgh, School of Engineering, Edinburgh, UK; C.Khirallah@ed.ac.uk
* Correspondence: dejanv@uns.ac.rs; Tel.: +381-21-485-2535

Abstract: Exponential increase in mobile video delivery will continue with the demand for higher resolution, multi-view and large-scale multicast video services. Novel fifth generation (5G) 3GPP New Radio (NR) standard will bring a number of new opportunities for optimizing video delivery across both 5G core and radio access network. One of the promising approaches for video quality adaptation, throughput enhancement and erasure protection is the use of packet-level random linear network coding (RLNC). In this work, we discuss the integration of RLNC into the 5G NR standard, building upon the ideas and opportunities identified in 4G LTE. We explicitly identify and discuss in detail novel 5G NR features that provide support for RLNC-based video delivery in 5G, thus pointing out to the promising avenues for future research.

Keywords: Random Linear Network Coding; Mobile Cellular Networks; 4G Long-Term Evolution (LTE); 5G New Radio (NR); Mobile video delivery

1. Introduction

Mobile video delivery continues its growth in volume and will reach estimated 78% of the total mobile data traffic by 2021, compared to 60% in 2016 [1]. In absolute values, during the same period (2016-2021), the total volume of mobile data traffic will experience seven-fold increase [1]. This increase is due to the combination of ever increasing resolutions of user handsets and proliferation of 4K/8K ultra high-definition (UHD) formats [2], fueled by evolution of innovative video services relying on multi-view and 360-degree video [3], enhanced broadcast [4] and peer-to-peer video services [5].

Evolution of mobile cellular infrastructure capable to cope with the surge of video traffic is necessity for meeting these predictions. Towards this end, 3rd Generation Partnership Project (3GPP) have just completed the first phase in defining a new fifth generation (5G) New Radio (NR) interface [6]. One of the resting pillars of 3GPP 5G NR is the support for enhanced Mobile Broadband (eMBB) services that target providing users with sufficient data rates to accommodate new, high-rate mobile video services. The support for eMBB will be achieved by novel throughput-enhancement solutions implemented both in radio access network (RAN) [7], such as massive multiple-input multiple-output (MIMO) antenna technology or migration to wider millimeter-wave bands (mmWave), and in core network (CN) domain, by supporting network slicing via network function virtualization (NFV) and software-defined networking (SDN) technologies [8].

For mobile video multicast/broadcast services, mobile cellular networks provide support in the form of 3GPP-defined enhanced Multimedia Broadcast/Multicast Service (eMBMS) [9]. However, majority of mobile video traffic represents over-the-top (OTT) video streaming such as progressive downloading (PD) and adaptive bitrate streaming (ABR) for which mobile cellular network protocols remain largely oblivious [10]. In order to optimize mobile video delivery, a number of solutions have been proposed in recent research studies, including video-aware resource allocation [11] and proactive edge caching and processing [12–14].

In this paper, we focus on random linear network coding (RLNC) as a scheme identified to create potentially high impact on flexible, efficient and reliable mobile video delivery. RLNC is a packet-level

erasure protection mechanism which is simple, efficient and has a number of useful features including rateless property, i.e., capability to produce arbitrary many encoded packets from a given source block, and network coding property, i.e., capability to increase throughput in certain network scenarios by re-encoding packets in intermediate network nodes [15–18]. We provide an overview of RLNC placed in the context of a packet-level data processing protocol sublayer that can be easily integrated at different layers of protocol stack within mobile video delivery environment. The RLNC sublayer can be further optimized and improved with respect to complexity and video quality using sparse and unequal error protection RLNC design [19–22]. In parallel with the review of RLNC, we provide an in-depth overview of mobile video delivery, focusing on unicast and multicast/broadcast mobile video services over fourth generation (4G) Long-Term Evolution (LTE) network [9,23–25]. We then move on to investigate how RLNC sublayer can be integrated as part of the 4G LTE mobile video delivery services, discussing various options for both video unicast and multicast/broadcast across different protocol layers. In the final part, we provide possible directions for optimizing mobile video delivery and integrating RLNC sublayer within upcoming 5G NR standard. In doing so, we explicitly identify and discuss in detail novel 5G NR features that could provide support for RLNC-based video delivery in 5G, thus pointing out to the promising avenues for future research.

The paper is organized as follows. In Section 2, we provide an overview of RLNC, presenting it as a basis of a modular RLNC-based protocol sublayer. We present performance measures and possible design extensions of the RLNC sublayer. In Section 3, we review in detail video delivery over 4G LTE mobile cellular networks, focusing on two main types of services: OTT video unicast and eMBMS-based video multicast/broadcast services. We present a review of practical mechanisms and academic investigations, targeting both CN and RAN design, that aim to support and enhance 4G LTE mobile video delivery. In Section 4, we discuss integration of RLNC sublayer across different layers of 4G LTE protocol stack. Based on these insights, in Section 5, we identify novel opportunities and challenges for RLNC sublayer and, in general, for mobile video delivery optimization, within 3GPP standardized 5G NR. The paper is concluded in Section 6.

2. Overview of Random Linear Network Coding

In this section, we provide a generic overview of a packet-level RLNC method for erasure protection across packet erasure channels in both the unicast and the multicast/broadcast scenario. We adopt a modular approach where RLNC is set as a core component of a RLNC protocol sublayer, whose integration in mobile video delivery solutions will be discussed in the rest of the paper. For more detailed account on the theory of RLNC, we refer interested reader to [15–18].

2.1. Introduction to RLNC

The system model under consideration contains RLNC encoder block at the transmitter and RLNC decoder block at one or more receivers, connected via independent packet erasure channels. An example with a single transmitter and a single receiver is illustrated in Fig. 1.

RLNC encoder block: The input to the RLNC encoder is a source block $\mathbf{s} = [\mathbf{s_1}, \mathbf{s}_2, \ldots, \mathbf{s}_K]$ containing K equal-length *source packets*, each containing L symbols of a finite field $\mathbb{F}_q$ of size q. RLNC encoder encodes $\mathbf{s}$ into a stream of *coded packets* $\mathbf{c} = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_N]$, where each encoded packet $\mathbf{c}_i$ represents a random linear combination of the source packets, i.e., $\mathbf{c}_i = \sum_{j=1}^{K} g_{i,j}\mathbf{s}_j$, and is of the same length L symbols of $\mathbb{F}_q$ as source packets. The *coding coefficients* $g_{i,j} \in \mathbb{F}_q$ are selected uniformly at random from $\mathbb{F}_q$, and for each $\mathbf{c}_i$, the associated set of coding coefficients forms the *coding vector* $\mathbf{g}_i = [g_{i,1}, g_{i,2}, \ldots, g_{i,K}]$. Note that the transmitter can produce arbitrarily many encoded packets N from K source packets in a rateless fashion. Fixing N, the set of encoded packets can be represented as $\mathbf{c} = \mathbf{s} \cdot \mathbf{G}^T$, where *coding matrix* $\mathbf{G}$ represents a $N \times K$ random matrix over $\mathbb{F}_q$. Frequently, systematic RLNC is also considered, where $\mathbf{c}_i = \mathbf{s}_i, 1 \leq i \leq K$, i.e., the first K coded packets are replicas of source packets, thus the first K rows of $\mathbf{G}$ represent $\mathbf{I}_{K\times K}$ identity matrix.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 February 2018 doi:10.20944/preprints201802.0099.v1

Peer-reviewed version available at Information 2018, 9, 72; doi:10.3390/info9040072

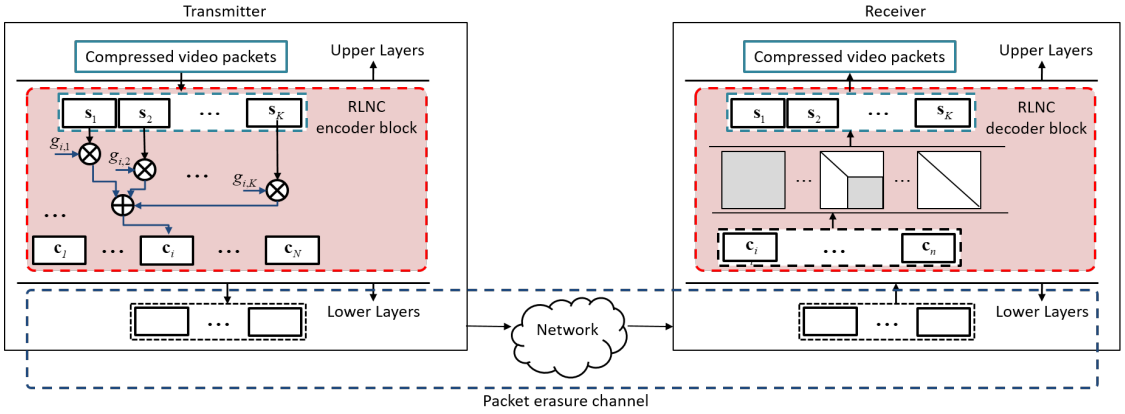

Figure 1. Generic RLNC sub-layers at the transmitter and the receiver side connected via packet erasure channel.

Packet erasure channel: Coded packets are transmitted to one or more receivers via independent packet erasure channels. For the j-th receiver, the erasure probability of the corresponding channel is denoted as ϵ_j, where $0 \leq \epsilon_j \leq 1$. If the focus is on a single receiver, we will omit index and use ϵ as erasure probability.

RLNC decoder block: Due to possible packet losses, a receiver receives a subset of $n \leq N$ coded packets. Extracting the corresponding coding vectors, the receiver obtains $\mathbf{c} = \mathbf{s} \cdot \mathbf{D}^T$, where, with some abuse of notation, $\mathbf{c} = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n]$ represents the set of received coded packets, while $\mathbf{D}$ is a random $n \times K$ matrix over $\mathbb{F}_q$. To recover source packets, the receiver applies Gaussian Elimination (GE) decoding, which successfully recovers the source block $\mathbf{s}$ iff for the rank $r(\mathbf{D})$ of the decoding matrix $\mathbf{D}$ it holds that $r(\mathbf{D}) = K$.

Performance measures: One of the main RLNC performance measures is the probability the source block $\mathbf{s}$ is successfully recovered given the number of received coded packets n. This *decoding probability* can be easily calculated [26][27]:

$$P_d(n) = \begin{cases} 0 & \text{if } n < K, \\ \prod_{i=1}^{K-1}(1 - \frac{1}{q^{n-j}}) & \text{if } n \geq K. \end{cases} \tag{1}$$

Note that (1) represents a cumulative distribution function (cdf) of the probability that K linearly independent packets are collected among n received coded packets. The corresponding probability density function (pdf) is $p_d(n) = P_d(n) - P_d(n-1)$.

It is often useful to take the transmitter perspective and introduce a time reference assuming that each coded packet transmission takes a unit-time slot. If we fix the number of transmitted coded packets to $N \geq K$, we can (re)define the decoding probability after N coded packets are transmitted:

$$P_d(N) = \sum_{n=K}^{N} \binom{N}{n} \epsilon^{N-n}(1-\epsilon)^n P_d(n). \tag{2}$$

Several interesting performance measures immediately follow. *Outage probability* $P_o(N)$ is a probability the receiver will not recover the source block after N transmitted coded packets: $P_o(N) = 1 - P_d(N)$. Average number of coded packet transmissions $\bar{N}$ required for successful source block decoding, also referred to as *average decoding delay*, can be investigated for fixed N, and for the case $N \to \infty$ [28]. For the former case, closed-form expressions for average decoding delay are available, while for the latter case, the upper bounds have been derived, both for the systematic and non-systematic RLNC [29].

2.2. Sparse RLNC

Standard RLNC described above is limited by the decoding complexity of the GE decoder that scales as $O(K^3)$ with the source block size. To some extent, and under low erasure rates, systematic RLNC approach may alleviate the complexity issue by removing received systematic packets from the decoding process. More flexible solutions resort to sparse RLNC (S-RLNC), where sparse random linear combinations are used to generate coded packets. This is typically done by changing the random sampling process of coding coefficients by promoting zero-valued coefficients:

$$\mathbb{P}(g_{i,j} = v) = \begin{cases} t & \text{if } v = 0, \\ \frac{1-t}{q-1} & \text{if } v \in \mathbb{F}_q \setminus \{0\}. \end{cases} \tag{3}$$

Similarly as in RLNC, S-RLNC schemes are also investigated for decoding probability, outage probability and average decoding delays. However, even for $P_d(N)$, exact expressions are unknown but only approximated by upper/lower bounds [19][20].

Adaptive extension to S-RLNC is proposed in the form of Tunable Sparse RLNC (TS-RLNC) [21]. In TS-RLNC, at the beginning of a session, sparse linear combinations are used, while as the session continues, the coding density is increased, which may be tuned for desired trade-off between decoding complexity and average decoding delay.

2.3. Unequal Error Protection RLNC

In typical video delivery scenarios, the source block $\mathbf{s}$ can be divided into L sub-blocks or *layers* such that $\mathbf{s} = [\mathbf{l_1}, \mathbf{l}_2, \ldots, \mathbf{l}_L]$, where the i-th layer $\mathbf{l}_i$ contains k_i source packets and $\sum_{i=1}^{L} k_i = K$. As the layer index i grows, the packets contained in $\mathbf{l}_i$ have progressively decreasing impact on reconstructed video quality. For this scenario, unequal error protection RLNC (UEP RLNC) schemes offer significant benefits in terms of flexibility, reconstructed video quality and decoding complexity, as compared to the standard RLNC.

Two generic and well-studied UEP RLNC methods are non-overlapping window RLNC (NOW-RLNC) and expanding window RLNC (EW-RLNC) [22]. In both schemes, a set of L windows $\mathbf{w} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_L]$ is defined over the set of layers: for NOW-RLNC, windows correspond to layers, i.e., $\mathbf{w}_i = \mathbf{l}_i$, while for EW-RLNC, the i-th window $\mathbf{w}_i$ contains the first i layers, i.e., $\mathbf{w}_i = [\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_i]$. Coded packets are produced by applying RLNC over a content of a window randomly selected using a window selection distribution $\mathcal{W}$. Proper design of $\mathcal{W}$ achieves desired balance of decoding probabilities of source packets belonging to different layers [22].

2.4. RLNC Extensions to Erasure Networks

In this paper, we restrict our attention to single-hop erasure channels, either in unicast or multicast/broadcast scenarios. Such models will be sufficient for our RLNC-based mobile video delivery considerations later in Sections IV and V. Before proceeding, we make two remarks.

RLNC and Rateless Codes: For single-hop channels considered in this paper, RLNC schemes described above represent instances of rateless codes [30]. Thus one can replace RLNC with other popular classes of rateless codes such as LT codes [31] or Raptor codes [32]. Indeed, Raptor codes provide near-optimal performance under significantly lower decoding complexity, thus allowing for larger source block lengths. The main benefit of using RLNC is that their usage is easily extended to multi-hop erasure network models where coding is performed in intermediate nodes. On the other hand, the source blocks lengths K in video delivery scenarios are typically small, thus using RLNC usually does not incur significant decoding complexity penalty.

Extensions to Erasure Network Models: RLNC emerged as a practical solution to the network coding problem, where RLNC is applied in intermediate nodes of erasure network models [17]. Since then, RLNC has been investigated in various erasure networks scenarios. Among these, we point out

to line networks [33], and more general multicast and multiple-unicast models [34][35], as the models of interest for RLNC applications in future dense mobile cellular networks.

3. Overview of Video Delivery in 4G Mobile Cellular Networks

This section reviews mobile video delivery in 4G LTE mobile cellular networks. We first provide background information on standard video content formats, unicast, and multicast/broadcast services in LTE. Then we provide specific details on LTE CN and RAN support for mobile video delivery.

3.1. Mobile Video Delivery in 4G LTE

Video Coding Standards: Video codecs are under constant evolution due to ever increasing performance requirements and novel use cases. The current video coding standards, H.265/HEVC [36], replaced the previous one, H.264/AVC [37], due to requirements for higher coding efficiency, higher spatial resolution (4K/8K video), color resolution and dynamic range. Extensions of HEVC include scalable (SHVC), multi-view (MV-HEVC), range (RExt) and 3D video coding (3D-HEVC) [38].

Details of HEVC compression are beyond the scope of this paper. We assume compressed video is packetized and organized into source blocks compatible with RLNC coding approach in Sec. II. Typically, source blocks represent compressed group of frames (GOFs). Layered source block structure can be obtained via scalable video coding (SVC) or using specific codec features such as slicing or data partitioning [23][39]. Finally, we note that, besides H.264/AVC and H.265/HEVC, other popular video codecs are in use such as VP9 and AV1 codecs.

Mobile Video Streaming/Downloading over 4G LTE: Most prevalent techniques for online video delivery are progressive downloading (PD) [24] and adaptive bitrate streaming (ABR) [25]. Chronologically, progressive downloading was first implemented and aimed to enable video users watching the video before the entire video content is downloaded. Video players download first metadata which describes video details and as soon as the first video data has been downloaded the rendering can start. ABR streaming also provides users capability to watch video content before the download is complete, however, this streaming technique provides multiple representations of the same video on the content server. These representations are encoded in different resolutions and bitrates thus allowing video clients to adapt delivered video resolution by switching between different representations according to the bandwidth available on the client side. Multiple ABR implementations are available but the most dominant ones are Apple HLS (HTTP Live Streaming), DASH (Dynamic adaptive streaming over HTTP), Microsoft Smooth Streaming and Adobe HTTP Dynamic streaming. Analysis shows that around 80% of total mobile streamed video is delivered in ABR format while the rest is delivered in PD format. 70% out of total mobile streamed video is delivered in encrypted format by using HTTPS or QUIC transport, where e.g., QUIC protocol is used for Youtube content.

Mobile Video Multicasting/Broadcasting over 4G LTE: LTE network support for a point-to-multipoint (PtM) services is defined in 3GPP as evolved Multimedia Broadcast/Multicast Service (eMBMS) [9]. The service is initially designed for mobile TV and radio broadcasting use case, however, other push-based services such as popular content caching (e.g., podcasts, news, ads, updates), live streaming of popular events (e.g., sport events such as olympics) and mobile network emergency alerts contributed to eMBMS development. eMBMS is delivered via PtM radio bearers thus providing for efficient usage of radio resources for the price of using fixed (i.e., non-adaptive) and conservative transmission configuration targeting improved cell coverage. In the context of this paper, eMBMS provides support for application-layer forward error correction (AL-FEC) [40], where Raptor codes are recommended, although potentially, RLNC could also be used as an alternative. Despite rising interest in LTE broadcasting, eMBMS has not yet been massively deployed at mobile network operators, while significant experience and promising prospects are gained via service trials [41].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 February 2018 doi:10.20944/preprints201802.0099.v1

Peer-reviewed version available at Information 2018, 9, 72; doi:10.3390/info9040072

3.2. Core Network Support for Mobile Video Delivery over 4G LTE

4G LTE Network Architecture: Figure 2 illustrates LTE network architecture that consists of two main parts: i) evolved universal terrestrial radio access network (E-UTRAN), and ii) evolved packet core (EPC). Due to functional split, EPC separates user plane and control plane elements. User plane elements, Serving-Gateway (S-GW) and Packet Data Network (PDN)-Gateways (P-GW), provide data connectivity between E-UTRAN and external PDN. The S-GW handles the user-plane packet data termination towards E-UTRAN, while P-GW interfaces with the external PDNs performing IP related functions such as IP address allocation, policy enforcement, packet classification and routing. The main control plane element, Mobility Management Entity (MME), is responsible for connection/release of radio bearers to user equipment (UE). Further control plane entities include the Policy and Charging Rules Function (PCRF), enforcing policies and rules that are configured statically or dynamically per subscriber data session, Home Subscriber Server (HSS) that contains subscriber data such as user/QoS/barring profiles, and Online Charging System (OCS) used for real time rating of subscriber data usage and providing subscribers with data usage control. For more details on EPC architecture and elements, we refer interested reader to [42][43].

Popular over-the-top (OTT) mobile video unicasting services such as ABR or PD do not require additional EPC elements, as they are transmitted transparently over the EPC data bearers. However, as described next, due to massive volume of ABR/PD traffic mobile operators often empower their EPC with video optimization (VO) platforms. In contrast, for video multicasting/broadcasting, significant EPC upgrade is needed to provide eMBMS, as presented in Fig. 2.

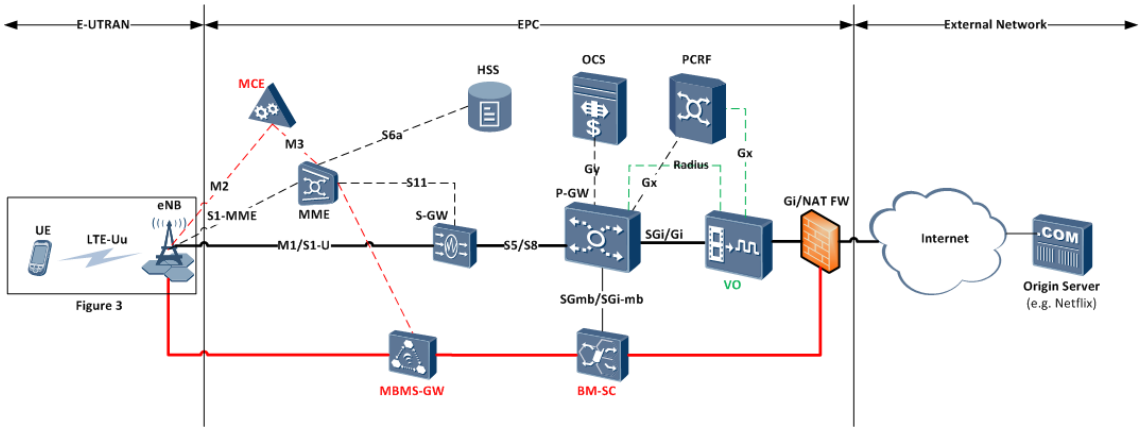

Figure 2. EPC network elements supporting mobile video delivery over 4G LTE.

Core Network Support for Video Streaming/Downloading: To provide support for optimized unicast video delivery, mobile operators introduce VO platforms (Fig. 2). VO platform is used to improve subscriber experience by optimizing video content to a format which will provide the best viewing experience for the given subscriber's network conditions. The VO platform consists of traffic classifier (TC) and the system that performs optimization of unicast streams (VO-subsystem). The video traffic, which is sent by content server towards a content consumer (mobile subscriber), is intercepted by the VO platform, where it is first detected by TC, passed to the VO-subsystem, optimized, and sent towards the content consumer. The most frequently used VO optimization methods are:

- **Transcoding:** converts video content from one format to another by changing e.g. encoding format, resolution, codecs, frame rate, etc. Online (on-the-fly) transcoding is mostly used. In addition, offline transcoding can be used and it is done in a way that some popular videos are downloaded and optimized in advance before being stored in cache [23][24].
- **Transrating:** converts video by keeping the original video format and resolution and by changing number of bits per pixel. This technique is not widely used [44].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 February 2018 doi:10.20944/preprints201802.0099.v1

Peer-reviewed version available at Information 2018, 9, 72; doi:10.3390/info9040072

Table 1. VO optimization methods and traffic types

Video Traffic Type	Possible Optimization Method
ABR over HTTP	ABR pacing; JIT pacing; ABR manifest file manipulation
ABR over HTTPS	ABR pacing; JIT pacing;
ABR over DRM over HTTP	ABR pacing; JIT pacing;
ABR over QUIC	ABR pacing; JIT pacing;
PD over HTTP	JIT pacing; online/offline transcoding; transrating; caching
PD over HTTPS	Not possible to optimize

- **ABR pacing:** receives traffic from a content server with one pace and send it towards a consumer with another pace in order to limit the representation quality requested by ABR clients on the subscriber side. This technique changes effective bandwidth perceived by ABR client side in order to affect the representation quality that will be selected by the client [44][45].
- **JIT (just-in-time) pacing:** receives traffic from a content server with one pace and lowers the downstream pace in "just in time" manner. This is done in order to avoid unnecessary filling of video player buffer on subscriber side as well as waste of network resources [24].
- **ABR manifest file manipulation:** consists of interception of ABR manifest file at the VO platform, parsing it and filtering out the representations from the manifest file that are not possible to reproduce on subscriber side under current network conditions.
- **Caching (transparent and selective):** consists of storing popular traffic on the platform in order to make it available for future subscriber's requests [46][47]. Transparent caching strategy consists of caching of all unencrypted content. As it is resource consuming operators typically uses selective caching such as caching of the content that is previously transcoded.

Table 1 provides the applicability of the above-listed techniques on different video traffic types. Whether the traffic is optimized by VO-subsystem depends on its profile configuration, which may depend on: i) time of the day (e.g., during peak hours more aggressive optimization can be done), ii) radio access technology (RAT) type of the subscriber (e.g., allowed max bitrate for 4G RAT type is higher than for 3G RAT type) which is received from P-GW via Radius interface (Fig. 2), iii) assigned subscriber policy received from PCRF via Gx interface (Fig. 2), iv) RAN congestion state (e.g., if RAN is in congested state then more aggressive optimization can be used for all subscribers to mitigate congestion), v) location of subscriber (e.g., if the subscriber is associated to congested cell then more aggressive optimization is applied), and vi) device class (e.g., streams destined to devices with small screens can be optimized in more aggressive way than streams to devices with a wide screens).

ABR over HTTPS is dominant video format, thus making ABR and JIT pacing the most commonly used optimization methods in VO platforms. We note that in academic studies, video quality based optimization methods for ABR video delivery is currently very popular research topic [45][48].

Core Network Support for Video Multicasting/Broadcasting: The entry point for eMBMS video content is broadcast/multicast service centre (BM-SC), which schedules and announces eMBMS services to end users. BM-SC is where AL-FEC is applied for packet-level erasure protection, which can be further optimized for eMBMS live streaming services [40][49]. Based on the content and control inputs from BM-SC, MBMS Gateway (MBMS-GW) establishes IP multicast session towards all eNBs that deliver eMBMS service to end users, supported by control signaling via MME. Single frequency network service (MBSFN) is commonly used, where the content is delivered with tight synchronization requirements enforcing identical physical layer (PHY) configuration at all eNBs to enhance reception at cell edges. Alternatively, single-cell point-to-multipoint service (SCPTM) can be used targeting eMBMS service in a specific cell [9]. An overview of network deployment options for both video unicast and eMBMS services over LTE are provided in [10].

As depicted in Fig. 2, the data path (solid red line) traverses BM-SC, MBMS-GW and proceeds via PTM IP Multicast session to the set of eNBs in MBSFN service area in the case of MBSFN service, or directly to a specific eNB in case of SCPTM service. The supporting signaling (dashed red line)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 February 2018 doi:10.20944/preprints201802.0099.v1

Peer-reviewed version available at Information 2018, 9, 72; doi:10.3390/info9040072

traverses MME and Multi-cell/Multicast Coordinating Entity (MCE), where MME is responsible for session-level management, while MCE controls configuration of eNBs in MBSFN area, such as coding and modulation, time synchronization and resource allocation [9]. Overall, providing eMBMS services requires significant CN upgrade, which, together with lack of killer applications [41], contributes for slow deployment of eMBMS at mobile operators.

3.3. Radio Access Network Support for Mobile Video Delivery over LTE

4G LTE Radio Access Network: RAN comprises large number of eNBs establishing radio connections to user devices (UEs). Figure 3 illustrates RAN user plane protocol stack for data delivery between eNB and UE. IP data flow towards UE passes through the Packet Data Conversion Protocol (PDCP) for header compression and ciphering, and Radio Link Control (RLC) protocol for segmentation/concatenation into suitably sized RLC packets that match the MAC frame size. If used in acknowledged automatic repeat request (ARQ) mode, RLC handles error-free and in-sequence RLC packet delivery between the eNB and UE. MAC layer introduces hybrid ARQ (HARQ) protection where, if MAC frame is not received correctly, up to three additional incremental redundancy MAC frames are transmitted. Finally, each MAC frame is allocated a single PHY transport block (PHY TB) placed into an OFDM-based time-frequency resource grid available to the eNB [43].

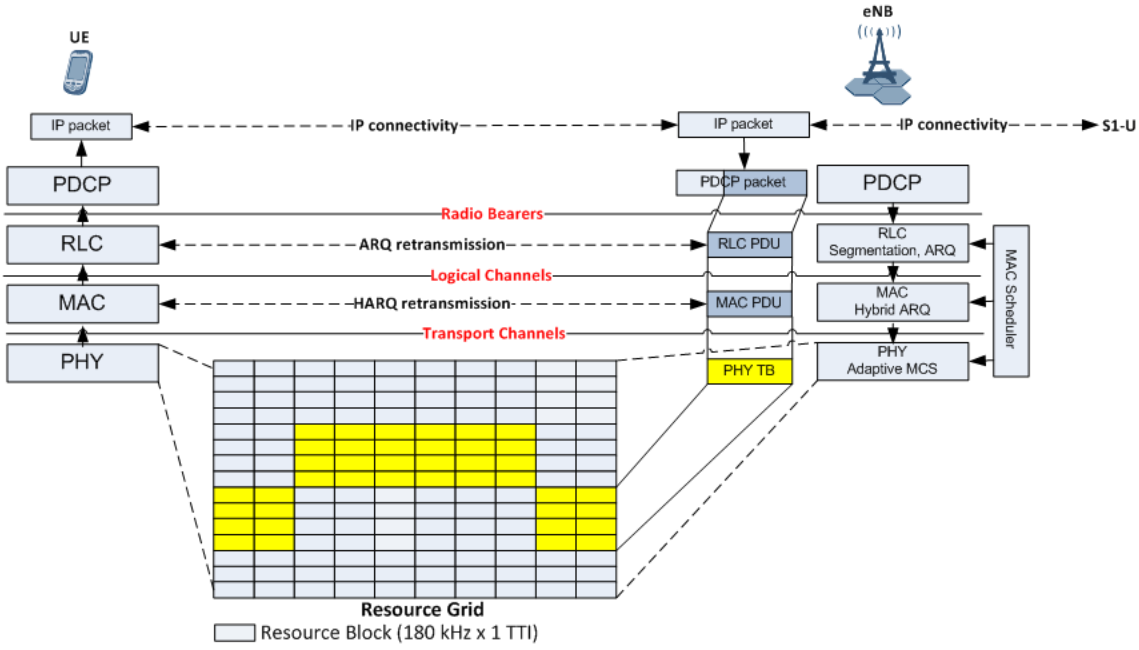

Figure 3. RAN network elements supporting mobile video delivery over 4G LTE.

PHY TBs of concurrent IP data flows are scheduled onto PHY resource blocks (PHY RBs) within every transmission time interval (TTI) of 1ms duration. A single PHY RB is a resource allocation unit of 1 TTI time duration and 12 OFDM carriers (180 kHz). The total number of PHY RBs per TTI depends on the bandwidth allocated to eNB (e.g., 50 PHY RBs for 10 MHz downlink channel). The information carrying capacity of PHY TBs depends on the number of PHY RBs allocated to the UE, adaptive modulation and coding scheme (MCS) applied and the multiple antenna (MIMO) mode used. Note that the user plane downlink and uplink data flows are multiplexed as part of the hierarchy of logical, transport and physical channels defined at LTE MAC and PHY layer. For example, at the lowermost layer, user data is carried by the physical downlink/uplink shared channel (PDSCH/PUSCH), along with a number of other physical control channels and channel reference signals. For a detailed overview of LTE RAN interface, we refer to more detailed exposition in [43].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 February 2018 doi:10.20944/preprints201802.0099.v1

Peer-reviewed version available at Information 2018, 9, 72; doi:10.3390/info9040072

Radio Access Network Support for Video Streaming/Downloading: eNB/UE interface represents the critical link in end-to-end video delivery from content servers to mobile devices, both in terms of channel capacity and variability. Optimization of resource allocation is a key factor for efficient usage of available radio spectrum. In LTE, MAC scheduler is responsible for allocation of PHY RBs to active UEs based on feedback on their channel quality indicators (CQI), and is usually based on proportional-fair (PF) scheduling [43][50]. However, standard MAC schedulers are typically oblivious to video traffic. As a tempting idea, large number of studies explored cross-layer optimized design with MAC schedulers directly using perceived video-quality information [11,51–53]. Although such a schemes offer performance gains, due to complexity of cross-layer implementation, they are rarely applied in practical systems.

Radio Access Network Support for Video Multicasting/Broadcasting: Besides CN upgrade, eMBMS requires additional radio resources for eMBMS service. Thus a new set of logical, transport and physical channels have to be configured at the RAN interface. For example, physical multicast channel (PMCH) is introduced to carry user plane eMBMS data to eMBMS users in the cell. Resource allocation, coding and modulation configuration at PMCH can be further optimized for efficient LTE video broadcasting [54–56], although these prospects are rarely applied at mobile operators.

4. Random Linear Network Coding for Mobile Video Delivery

In this section, we bring together the material presented in previous two sections and discuss integration of RLNC sublayer across different layers of 4G LTE mobile video delivery environment.

4.1. RLNC: Where should it be?

RLNC represents a flexible packet-level erasure coding sublayer that can be easily integrated at different positions within the protocol stack. RLNC flexibility rests on flexible definition of input source blocks, source block length K, source/encoded packet length L, coding coefficient field size q, number of encoded packets N to be produced, and other RLNC properties such as sparsity, tunability and UEP. Furthermore, simple analytical expressions and bounds for packet decoding probabilities, outage probabilities and average decoding delays provide for optimized RLNC design in different scenarios, as detailed in Sec. 2. In the following, we discuss the suitable position for RLNC sublayer in the context of mobile video delivery. Possible opportunities for RLNC sublayer placement are divided into: 1) End-to-end solutions for RLNC residing either at application or transport layer, and 2) RLNC solutions residing within RAN protocol stack, as illustrated in Figure 4.

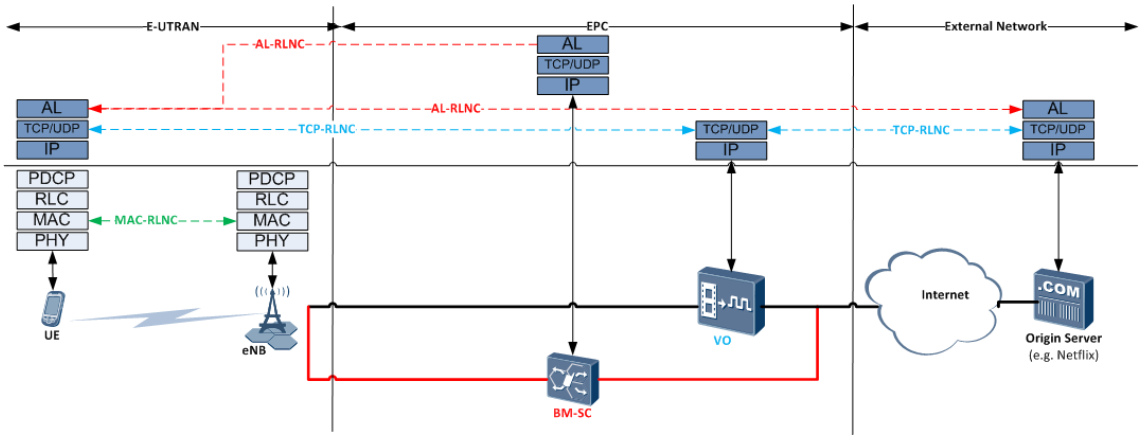

Figure 4. RLNC sublayer position options for mobile video delivery.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 February 2018 doi:10.20944/preprints201802.0099.v1

Peer-reviewed version available at Information 2018, 9, 72; doi:10.3390/info9040072

4.1.1. End-to-end Solutions for RLNC

Application Layer RLNC: The simplest and easiest approach to integrate RLNC sublayer into mobile video delivery is to perform application-layer RLNC (AL-RLNC). In particular, for multicast/broadcast scenario (eMBMS), the required infrastructure is already in place, as eMBMS provides native support for AL-FEC [9]. However, in this case, RLNC provides similar performance as compared to already proposed AL-FEC solutions such as Raptor codes [57] or LDPC triangle/staircase codes [58], providing limited benefit since no re-encoding at intermediate nodes is typically assumed in multicast/broadcast setup. Nevertheless, a large number of academic studies considered application and optimization of RLNC for mobile multicast video delivery demonstrating various benefits in different scenarios [59–62]. Besides notable benefits of AL-RLNC in terms of flexibility and ease of implementation, there are several drawbacks of AL-RLNC worth emphasizing. AL-RLNC introduces redundancy either at the content server or BM-SC (eMBMS) thus adding significant communication overhead across the entire end-to-end IP multicast session, including typically reliable and over-provisioned core network optical links. In addition, AL-RLNC is transparent to lower layers and, in practice, it is hard to provide cross-layer optimized solution, e.g., at unreliable eNB/UE interface, that takes into account AL-RLNC. For unicast video streaming services, AL-RLNC is usually not considered as the most popular ABR/PD streaming solutions rely on HTTP via TCP, which already provides reliable packet delivery. However, this might change as more and more ABR traffic moves to Quick UDP Internet Connections (QUIC) protocol which relies on UDP [63]. However, note that most of the current unicast video traffic is encrypted, which might introduce practical limitations in applying RLNC at all layers below the application layer.

Transport Layer RLNC: Another possibility for end-to-end RLNC is integration of RLNC sublayer in transport layer protocols. In the case of TCP, a groundbreaking idea has been investigated in which, instead of source packets, TCP transmits coded packets, and where instead of acknowledgement of individual source packets, the TCP receiver acknowledges received "degrees of freedom", i.e., the rank of the decoding matrix currently available at the receiver [64][65]. Clearly, coded TCP is suitable solution for unicast ABR/PD video services which rely on TCP, thus any throughput improvements of coded TCP over traditional TCP would reflect on OTT video traffic. In addition, with VO platform in-the-middle (Sec. 3.2), it is possible to split TCP connection between the ABR/PD streaming client and the content server, thus independently optimizing each of the two resulting TCP connections: the one between the UE and the VO platform, and the one between the VO platform and the content server. As a potential obstacle to coded TCP, its impact on TCP congestion control has yet to be better understood [66].

4.1.2. RLNC Solutions in Radio Access Network

As an alternative to integration at the higher layers, RLNC sublayer could be integrated where the network reliability is critical: within RAN protocols. This solution could be equally applied for unicast and multicast/broadcast mobile video delivery, as it is triggered over the packets delivered between eNB and UE(s) either via unicast (PDSCH) or multicast (MTCH) transport channels.

The solution for RLNC within the MAC layer of LTE RAN protocol stack has been investigated in detail in [67]. The proposed MAC-RLNC solution, integrated as the upper sublayer within MAC protocol, adopts the built-in flexibility of RLC protocol (segmentation and concatenation) to define the desired size of the source block that will be provided to the MAC-RLNC sublayer in the form of RLC packet data unit (PDU). From each received RLC PDU, MAC-RLNC produces a stream of carefully-sized fixed-length coded packet which are encapsulated in consecutive MAC frames and delivered via underlying PHY TB containers to the receiver side. The rationale behind MAC-RLNC was to apply rateless RLNC concept and convert RLC PDU transmission into a "fluid" delivery of coded packets flexible to fit PHY TB containers of different sizes. This way, MAC-RLNC would essentially replace MAC-based HARQ protocol, as instead of sending HARQ retransmissions, the transmitter simply continues to send a new set of coded packets, until receiver acknowledges it has received

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 February 2018 doi:10.20944/preprints201802.0099.v1

Peer-reviewed version available at Information 2018, 9, 72; doi:10.3390/info9040072

full-rank set and is able to recover the RLC PDU (see Figure 5). Performance of such a MAC-RLNC scheme has been investigated in terms of video delivery over LTE [68]. Integration of RLNC into RAN protocols opens a novel possibilities for joint optimization of RLNC and resource allocation for both unicast and multicast/broadcast services, the topic we present in more details in the following subsection.

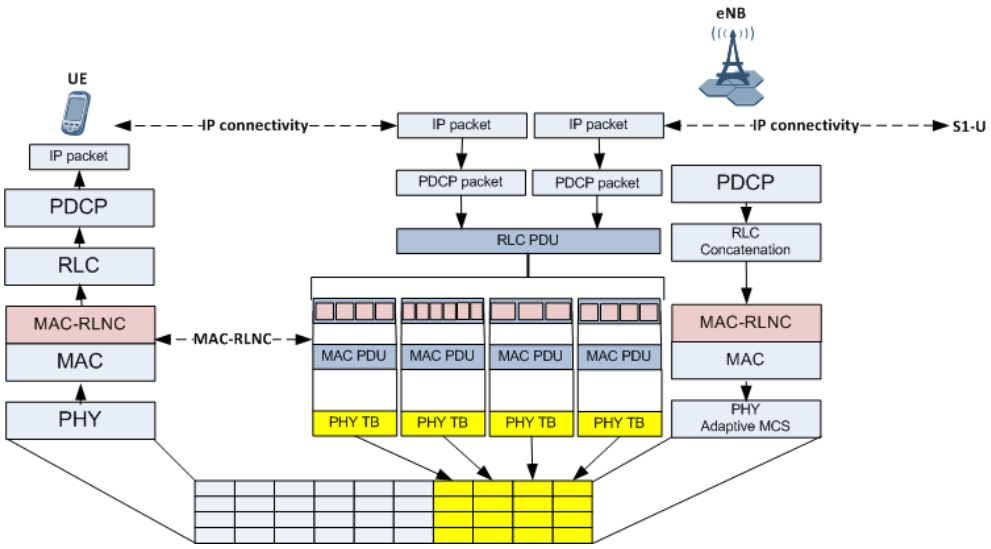

Figure 5. MAC-RLNC sublayer as part of LTE RAN protocol stack.

4.2. RLNC and Resource Allocation

Integration of RLNC sublayer within LTE protocol stack opens novel and interesting problems of interaction between RLNC and resource allocation that we describe next.

We consider a general system model where a base station broadcasts an ℓ-layer video stream encoded according to the SVC paradigm. We also say that layer 1 is the basic and layers $2, \ldots, L$ are the enhancement layers. Assuming our system model adopts a NOW-RLNC implementation, any resource allocation strategy has to answer the following questions [19,69,70,72]:

- How many coded packet transmissions per-video layer are to be scheduled for transmission?
- What MCSs are to be used for broadcasting each coded packet?

Obviously, choosing the correct number of coded packet transmission affects the user decoding probability. Likewise, the MCS selection has a direct impact on the total number of users capable of successfully decoding a given number of video layers.

Let $\mathcal{F}$ be our utility function, a General Resource Allocation Problem (GRAP) for network coded video application has the following structure:

$$\text{(GRAP)} \quad \min_{\mathbf{m},\mathbf{N}} \ \mathcal{F} \tag{4}$$

$$\text{subject to} \quad m_{\ell-1} < m_{\ell}, \qquad \ell = 2, \ldots, L \tag{5}$$

$$U_{\ell} \geq \hat{U}_{\ell}, \qquad \ell = 1, \ldots, L \tag{6}$$

$$\mathcal{S}(\mathbf{N}) \leq \hat{S} \tag{7}$$

$$\mathcal{T}(\mathbf{N}) \leq \hat{T} \tag{8}$$

where m_{ℓ} signifies the MCS to be used for transmitting coded packets associated with layer ℓ. In addition, we define our optimization variables to be the vectors $\mathbf{m} = \{m_1, \ldots, m_L\}$ and $\mathbf{N} = \{N_1, \ldots, N_L\}$. Constraint (5) ensures that coded packets associated with layer $\ell - 1$ are being

transmitted with a smaller MCS compared to that used for layer ℓ – thus ensuring coded packets associated with layer $\ell - 1$ are received with a PER ϵ that is smaller than or equal to that experienced in receiving coded packets associated with layer ℓ. From (2), it follows that the probability of a user successfully recovering the first ℓ video layers can be expressed as $\prod_{i=1}^{\ell} P_d^i(N_i)$, where N_i is the number of coded packet transmission associated with layer i. Let us signify with U_ℓ, the number of users that can successfully decode the first ℓ video layers with a probability equal to or greater than $\hat{p}_d$. Constraint (6) ensures that U_ℓ is greater than or equal to a target number of users $\hat{U}_\ell$. We observe that relation $\hat{U}_{\ell-1} \geq \hat{U}_\ell$ holds true for $\ell = 2, \ldots, L$ as it would be pointless to impose video layer ℓ to be decoded by a number of users larger than $\ell - 1$. Constraint (7) ensures that the total number of coded packets scheduled on each radio frame does not exceed a given threshold $\hat{S}$. Finally, constraint (8) ensure that each portion of video stream is transmitted by a time $\hat{T}$.

Unfortunately, to the best of our knowledge, GRAP is an integer (potentially) non-linear optimization problem with no obvious analytical solutions. However, when $\mathcal{F}$ is equal to the total number of coded packet transmissions, authors' [69,70] argued that it is possible to find a good-quality feasible solution to GRAP using a two-step heuristic operating as follows:

1. *MCS Allocation* – The heuristic iterates over all the MCSs (starting from the highest), and for each video layer, it identifies the largest MCS such that constraints (5) and (6) are met. The output of this heuristic step is an instance of vector $\mathbf{m}$.
2. *Code Packet Transmissions Optimisation* – On the basis of the instance of vector $\mathbf{m}$ determined in the previous step, the minimum value of N_ℓ is determined, for any $\ell = 1, \ldots, L$ - thus an instance of vector $\mathbf{N}$ is found.

It is easy to prove the aforementioned heuristic has a reduced computational complexity and when it identifies an instance of $\mathbf{m}$ and $\mathbf{N}$ they jointly constitute a feasible solution to GRAP. Since $\mathbf{m}$ and $\mathbf{N}$ are independently optimized, the heuristic solution is not always optimum. However, for realistic network and video service deployments, the heuristic provides good-quality solutions [70,71].

It has also been observed that GRAP can be extended to system models where EW-RLNC implementations are in use [70]. In these cases, terms $N_1, \ldots, N_L$ represents the number of coded packet transmissions associated with expanding windows $\mathbf{w_1}, \ldots, \mathbf{w_L}$ to be scheduled for transmission. Similarly, terms $m_1, \ldots, m_L$ identifies the MCSs to be used for transmissions associated with $\mathbf{w_1}, \ldots, \mathbf{w_L}$, respectively. As observed in Section 2.3, a user can decode the first ℓ video layers if it successfully recovers any of the expanding windows $\mathbf{w}_\ell, \mathbf{w}_\ell + \mathbf{1}, \ldots, \mathbf{w_L}$. As such, in this case, the probability of a user successfully recovering the first ℓ video layers is $\bigvee_{i=\ell}^{L} P_{d,\mathbf{w_i}}(\mathbf{N})$, where $P_{d,\mathbf{w_i}}(\mathbf{N})$ is the probability of a user recovering $\mathbf{w_i}$ – thus terms $\mathcal{U}_1, \ldots, \mathcal{U}_L$ have to be redefined accordingly. Despite its new setting, the aforementioned two-step heuristic can still be applied to find good-quality feasible solution to GRAP [70].

5. Random Linear Network Coding for 5G New Radio

Standardization of 5G NR technology is currently under way within 3GPP. In the moment of writing of this paper, the Phase 1 of the 5G NR standardization has just been completed [6]. In this section, we provide a short intro to 5G NR, followed by identification of several 5G NR standard features suitable for further study of RLNC integration and optimization of mobile video delivery in 5G. We conclude the section with recent related work on video delivery deployment studies in 5G.

5.1. Introduction to 5G NR

Overall 5G NR Architecture: The 5G system architecture consists of a 5G Access Network (AN), 5G Core (5GC) Network and the UE [73]. The 5G AN comprises an NG-RAN and/or non-3GPP AN connecting to a 5G Core Network. NG-RAN focuses on the radio interface protocol architecture and contains NG-RAN nodes termed next-generation NodeB (gNB), providing NR user plane and control plane protocol terminations towards the UE. The gNBs are interconnected with each other and are

Peer-reviewed version available at Information 2018, 9, 72; doi:10.3390/info9040072

also connected by means of the NG interfaces to the 5GC, most importantly to the AMF (Access and Mobility Management Function) and to the UPF (User Plane Function), as illustrated in Figure 6 [74]. For the details on 5G NR architecture and other 5GC entities and interfaces, we refer the interested reader to [73].

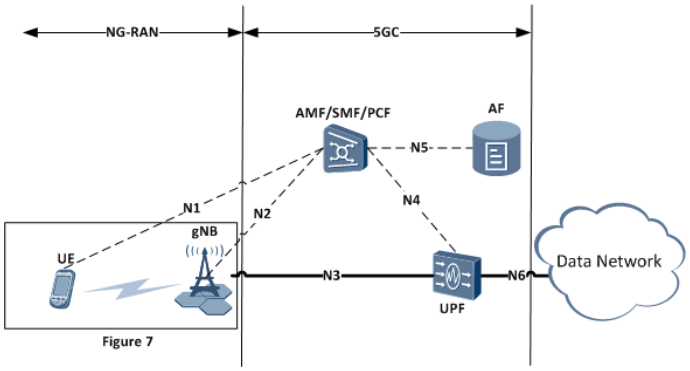

Figure 6. 5G NR System Architecture.

5G NR Radio Protocol Architecture - User Plane: Figure 7 (left) illustrates the protocol stack for the gNB user plane that includes Service Data Adaptation Protocol (SDAP), PDCP, RLC, MAC and PHY sublayers. The new SDAP sublayer is introduced to 5G NR targeting specifically Quality of Service (QoS) control in NG-RAN. SDAP handles: i) the mapping of 5G QoS flows to data radio bearers (DRBs), and ii) marking of 5G QoS flow Identifier (QFI) to a QoS flow in both DL and UL packets. We will discuss in more detail 5G QoS architecture later in this subsection.

The main functions of PDCP sublayer remain similar to 4G LTE (e.g., header compression, ciphering and integrity protection), with the exception of one important new feature called PDCP PDU duplication. The PDCP duplication offers the possibility of sending the same PDCP PDUs twice: once on the original RLC entity and a second time on the additional RLC entity. The original and duplicate PDCP PDUs are transmitted on different carriers. The two different logical channels can either belong to the same MAC entity, e.g., via carrier aggregation (CA), or to different logical channels in the case of multi-connectivity, i.e., dual connectivity (DC) [74]. As we discuss later, combined with the PDCP duplication feature, PDCP could provide an ideal place for RLNC sublayer integration.

The RLC protocol preserved most of the LTE functionalities, including ARQ-based error correction, segmentation and reassembly of SDUs, etc. Similarly, the main services and functions of the MAC sublayer resemble those in LTE and include, e.g., error correction through HARQ, priority handling between UEs by means of dynamic scheduling, and priority handling between logical channels of one UE by means of logical channel prioritisation.

QoS Architecture in 5G NR: Basic granularity for QoS control in LTE EPC/E-UTRAN is EPS bearer/E-UTRAN Radio Access Bearer (E-RAB). EPS bearer/E-RAB established when UE connects to a PDN is called the default bearer, while any additional bearer is referred to as a dedicated bearer. Each bearer is characterized by the same packet forwarding treatment (e.g., scheduling policy, queue management policy, rate shaping policy, RLC configuration, etc.). A bearer is called guaranteed bit-rate (GBR) bearer if it is provided dedicated network resources, otherwise, it is called non-GBR bearer.

In 5G, the concept of QoS flow is introduced which is considered the finest data flow granularity that receives the same forwarding treatment. Providing different QoS forwarding treatment requires separate 5G QoS flows. Multiple QoS flows are part of a packet data unit (PDU) session: an association between the UE and a PDN [73]. For each UE, the 5GC may establish one or more PDU sessions carrying QoS flows of different QoS level. QoS flow ID (QFI) identifies a QoS flow within a PDU session and indicates a specific QoS forwarding behavior (e.g. packet loss rate, packet delay budget). At the NG-RAN, the QoS flows are mapped to one or more data radio bearers (DRBs). NG-RAN and 5GC jointly ensure QoS by mapping packets to appropriate QoS flows and DRBs by a 2-step mapping

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 February 2018 doi:10.20944/preprints201802.0099.v1

Peer-reviewed version available at Information 2018, 9, 72; doi:10.3390/info9040072

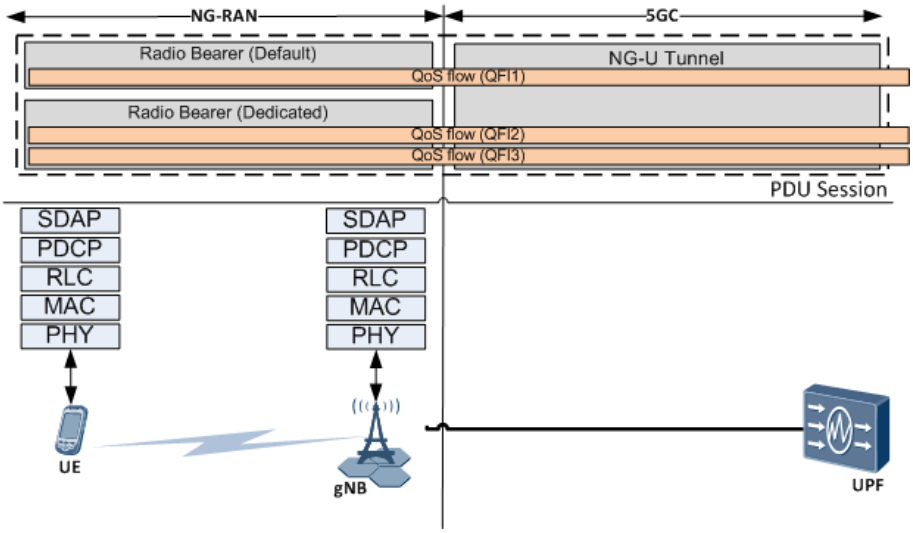

Figure 7. 5G NG-RAN Protocol Stack and 5G QoS Architecture.

process, where an IP-flow is mapped by 5GC to QoS flows while NG-RAN (SDAP protocol) maps QoS flows to DRBs. Within each PDU session, NG-RAN (SDAP) decides how to map QoS flows to DRBs. The QoS architecture in 5G is illustrated in Figure 7.

5.2. Opportunities and Challenges for Mobile Video Delivery in 5G

PDCP Coded Duplication in 5G NR: Due to PDCP duplication feature, PDCP protocol becomes an interesting option for the RLNC sublayer within the NG-RAN user plane protocol stack. Empowering PDCP layer with RLNC sublayer could provide a possibility for "coded duplication" where, instead of a simple duplication of PDCP PDUs, one could transmit two different sets of RLNC coded packets created from the appropriately defined source block. Next, we briefly describe the concept of coded duplication but leave the more detailed investigation for our future work.

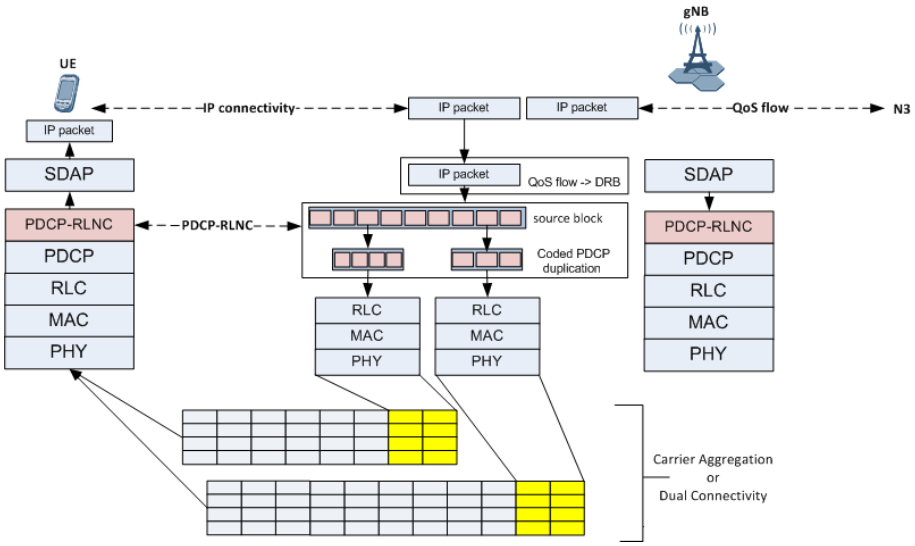

Figure 8. The concept of Coded PDCP Duplication in 5G NR.

Figure 8 illustrates the concept of PDCP Coded Duplication. PDCP protocol receives PDCP service data units (SDUs), i.e., IP packets, and organizes one or more PDCP SDUs into a source block of length K source packets. The source block is processed by RLNC sublayer to produce N coded packets

Peer-reviewed version available at Information 2018, 9, 72; doi:10.3390/info9040072

encapsulated into a sequence of coded PDCP PDUs. The set of PDCP PDUs carrying the coded packets of the same source block are enumerated using the same PDCP sequence number and transmitted using PDCP duplication feature via separate protocol legs. The coded duplicated content propagates down the two independent legs, undergoing possibly different segmentation at the two RLC entities followed by different MAC frame delivery success. Note that due to RLNC coding at PDPC layer, RLC layer ARQ and MAC layer HARQ could be altogether avoided in coded PDCP duplication.

At the receiving side, instead of removal of PDCP duplicates, the receiving PDCP entity would collect the coded content from all incoming PDCP PDUs tagged with the same sequence number and received through both legs. The coded packets are extracted from PDCP PDU until the content of the source block is reconstructed, as illustrated in Figure 8. Finally, note that coded PDCP duplication operates across two parallel packet erasure channels, which offers different possibilities for matching layered UEP RLNC schemes and parallel erasure channels, empowered with appropriate resource allocation of coded packets to coded duplicated PDCP PDUs.

QoS Control for 5G Mobile Video Delivery: The 5G QoS architecture described in the previous subsection will provide advanced mechanisms for QoS control in mobile video delivery. 5GC will classify QoS flows based on their QFI and provide per flow forwarding treatment in terms of e.g. packet loss rates and packet delay, thus providing mobile video flows with desired delivery parameters.

One of the key elements of 5G QoS architecture is a novel SDAP protocol defined at NG-RAN user plane interface. SDAP is additional resource allocation entity in 5G NR that will complement MAC scheduler. While the "lower layer" MAC scheduler aims to dynamically schedule resource blocks to different UEs in order to maintain guaranteed or best possible DRB parameters, the "upper layer" SDAP scheduler aims to assign QoS flows to different DRBs in order to satisfy their QFI-defined parameters. Note that different DRBs defined at the interface towards the UE may be configured using different NG-RAN protocol configurations. In this sense, coded PDCP duplication described above may provide a flexible approach to trade off reliability and latency and fine tune QFI-defined requirements in terms of packet loss rates and packet delay for a given DRB.

5.3. Related Studies on Mobile Video Delivery in 5G NR

Without the goal of being exhaustive, we finalize this review paper with an overview of related work in the domain of mobile video delivery in 5G. Along with our discussion above, these papers could help the reader to identify possible research avenues in the domain of 5G mobile video delivery.

Starting with the core network aspects, mobile edge computing (MEC) and edge content caching are identified as promising NG-RAN architectures that will greatly improve massive unicast OTT video delivery services such as ABR. In [75], the combination of network function virtualization (NFV) and MEC is explored for deployment of context-aware virtualized adaptive prefetching agents at the mobile edge that will provide QoS-guaranteed UHD video services. In ultra-dense 5G network scenario, the work presented in [77] advocates local caching of video in small cells in combination with device-to-device (D2D) communication, while in [76], the authors consider network-aware ABR video content caching at the network edge combined with end-to-end video streaming resource allocation optimization. Edge caching is jointly optimized with multicast transmissions in what authors term as multicast-aware edge caching for 5G video delivery of popular content in [78]. We also point to the multi-server video streaming architecture for optimized ABR video delivery in 5G, relying on a combination of cloud RAN (C-RAN) and MEC concepts, as detailed in [79].

In RAN domain, we emphasize several studies related to the work presented here. The work in [80] presents a study on high-quality high-throughput video streaming with enhanced reliability, which is achieved using a combination of end-to-end RLNC and 5G multi-connectivity via legacy LTE and 5G mmWave radio links. Our discussion on coded PDCP duplication draws inspiration from the recent work on optimized interface diversity for 5G ultra-reliable and low-latency services (URLLC) [81]. Both studies point out to the potential of using multi-connectivity (e.g., via carrier aggregation or dual-connectivity) for reliable and low-delay video delivery in 5G. Finally, we also emphasize the

role of resource management in NG-RAN, as recently explored in 5G mobile vehicular video delivery in [82]. Resource management in NR-RAN will require novel ideas for optimized cross-layer video optimization exploiting both SDAP-based allocation of QoS flows to DRBs and MAC-based dynamic scheduling of radio resources to UE terminals.

6. Conclusions

The goal of this paper was to present a detailed study of two interrelated topics, RLNC for packet erasure protection and mobile video delivery in mobile cellular systems. We introduced RLNC using module-based approach, motivated by the quest to identify both the need and the suitable location for RLNC sublayer in video delivery solutions for 4G/5G mobile cellular networks. Evolving from RLNC sublayer integration in 4G LTE, the paper culminates with investigation of 5G NR architecture and possible RLNC integration therein for future 5G optimized mobile video delivery. Across the paper, we used the opportunity to provide detailed review and point towards relevant publications of all the fundamental concepts related to mobile video delivery in 4G/5G networks.

References

1. Cisco Visual Networking Index: https://www.cisco.com/c/en/us/solutions/service-provider/visual-networking-index-vni
2. Ye, Y. and Andrivon, P., "The scalable extensions of HEVC for ultra-high-definition video delivery," *IEEE MultiMedia,* 21(3), pp.58-64, 2014.
3. Bastug, E., Bennis, M., Médard, M. and Debbah, M., "Toward interconnected virtual reality: Opportunities, challenges, and enablers," *IEEE Communications Magazine,* 55(6), pp.110-117, 2017.
4. de la Fuente, A., Leal, R.P. and Armada, A.G., "New technologies and trends for next generation mobile broadcasting services," *IEEE Communications Magazine,* 54(11), pp. 217-223, 2016.
5. Xu, C., Jia, S., Zhong, L. and Muntean, G.M., "Socially aware mobile peer-to-peer communications for community multimedia streaming services," *IEEE Communications Magazine,* 53(10), pp.150-156, 2015.
6. 3GPP 5G NR Rel. 15 Specification Series, http://www.3gpp.org/DynaReport/38-series.htm
7. Marsch, P., Da Silva, I., Bulakci, O., Tesanovic, M., El Ayoubi, S.E., Rosowski, T., Kaloxylos, A. and Boldi, M., "5G radio access network architecture: Design guidelines and key considerations,' IEEE Communications Magazine, 54(11), pp.24-32, 2016.
8. Rost, P., Banchs, A., Berberana, I., Breitbach, M., Doll, M., Droste, H., Mannweiler, C., Puente, M.A., Samdanis, K. and Sayadi, B., "Mobile network architecture evolution toward 5G," *IEEE Communications Magazine,* 54(5), pp.84-91, 2016.
9. Lecompte, D. and Gabin, F., "Evolved multimedia broadcast/multicast service (eMBMS) in LTE-advanced: overview and Rel-11 enhancements," *IEEE Communications Magazine,* 50(11), pp. 68-74, 2012.
10. Oyman, O., Foerster, J., Tcha, Y.J. and Lee, S.C., "Toward enhanced mobile video services over WiMAX and LTE," *IEEE Communications Magazine,* 48(8), 2011.
11. Luo, H., Ci, S., Wu, D., Wu, J., Tang, H, "Quality-driven cross-layer optimized video delivery over LTE," *IEEE Communications Magazine,* 48(2), 2010.
12. Wang, X., Chen, M., Taleb, T., Ksentini, A. and Leung, V., "Cache in the air: exploiting content caching and delivery techniques for 5G systems," *IEEE Communications Magazine,* 52(2), pp.131-139, 2014.
13. Bastug, E., Bennis, M. and Debbah, M., "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Communications Magazine,* 52(8), pp.82-89, 2014.
14. Tran, T.X., Hajisami, A., Pandey, P. and Pompili, D., "Collaborative mobile edge computing in 5G networks: New paradigms, scenarios, and challenges," *IEEE Communications Magazine,* 55(4), pp.54-61, 2017.
15. Ho, T., Mèdard, M., Koetter, R., Karger, D.R., Effros, M., Shi, J. and Leong, B., "A random linear network coding approach to multicast," *IEEE Transactions on Information Theory,* 52(10), pp.4413-4430, 2006.
16. Fragouli, C., Le Boudec, J.Y. and Widmer, J., "Network coding: an instant primer," *ACM SIGCOMM Computer Communication Review,* 36(1), pp.63-68, 2006.
17. Chou, P.A. and Wu, Y., "Network coding for the internet and wireless networks," *IEEE Signal Processing Magazine,* 24(5), pp.77-85, 2007.

18. Fragouli, C. and Soljanin, E., "Network coding fundamentals," *Foundations and Trends in Networking,* 2(1), pp.1-133, 2007.
19. Tassi, A., Chatzigeorgiou, I. and Lucani, D.E., "Analysis and optimization of sparse random linear network coding for reliable multicast services," *IEEE Transactions on Communications,* 64(1), pp.285-299, 2016.
20. Brown, S., Johnson, O. and Tassi, A., "Reliability of Broadcast Communications Under Sparse Random Linear Network Coding," to appear, *IEEE Transactions on Vehicular Technology,* 2018.
21. Feizi, S., Lucani, D.E., Sørensen, C.W., Makhdoumi, A. and Mèdard, M., "Tunable sparse network coding for multicast networks," *Network Coding (NetCod) 2014* pp. 1-6), 2014.
22. Vukobratovic, D. and Stankovic, V., "Unequal error protection random linear coding strategies for erasure channels," *IEEE Transactions on Communications,* 60(5), pp.1243-1252, 2012.
23. Schierl, T., Stockhammer, T., Wiegand, T., "Mobile video transmission using scalable video coding," *IEEE transactions on circuits and systems for video technology,* 17(9), pp. 1204-1217, 2007.
24. Ma, K.J., Bartos, R., Bhatia, S. and Nair, R., "Mobile video delivery with HTTP," *IEEE Communications Magazine,* 49(4), 2011.
25. Oyman, O. and Singh, S., "Quality of experience for HTTP adaptive streaming services," *IEEE Communications Magazine,* 50(4), 2012.
26. Cooper, C., "On the distribution of rank of a random matrix over a finite field," *Random Structures and Algorithms,* 17(3-4), pp.197-212, 2000.
27. Trullols-Cruces, O., Barcelo-Ordinas, J.M. and Fiore, M., "Exact decoding probability under random linear network coding," *IEEE Communications Letters,* 15(1), pp.67-69, 2011.
28. Nistor, M., Lucani, D.E., Vinhoza, T.T., Costa, R.A. and Barros, J., "On the delay distribution of random linear network coding," *IEEE Journal on Selected Areas in Communications,* 29(5), pp.1084-1093, 2011.
29. Chatzigeorgiou, I. and Tassi, A., "Decoding delay performance of random linear network coding for broadcast," *IEEE Transactions on Vehicular Technology,* 66(8), pp.7050-7060, 2017.
30. Liva, G., Paolini, E. and Chiani, M., "Performance versus overhead for fountain codes over $\mathbb{F}_q$," *IEEE Communications Letters,* 14(2), pp. 178-180.
31. Luby, M., "LT codes," *Foundations of Computer Science FOCS 2002,* pp. 271-280, 2002.
32. Shokrollahi, A., "Raptor codes," *IEEE Transactions on Information Theory,* 52(6), pp.2551-2567, 2006.
33. Pakzad, P., Fragouli, C. and Shokrollahi, A., "Coding schemes for line networks," *IEEE Int'l Symp. Information Theory ISIT 2005,* pp. 1853-1857, 2005.
34. Lun, D.S., Mèdard, M. and Koetter, R., "Network coding for efficient wireless unicast," *Int'l Zurich Seminar on Communications,* pp. 74-77, 2006.
35. Seferoglu, H., Markopoulou, A. and Ramakrishnan, K.K., "I^2NC: Intra-and inter-session network coding for unicast flows in wireless networks," *IEEE INFOCOM 2011,* pp. 1035-1043, 2011.
36. Sullivan, G.J., Ohm, J., Han, W.J. and Wiegand, T., " Overview of the high efficiency video coding (HEVC) standard," *IEEE Transactions on circuits and systems for video technology,* 22(12), pp.1649-1668, 2012.
37. Wiegand, T., Sullivan, G.J., Bjontegaard, G. and Luthra, A., "Overview of the H. 264/AVC video coding standard," *IEEE Transactions on circuits and systems for video technology,* 13(7), pp.560-576, 2007.
38. Sullivan, G.J., Boyce, J.M., Chen, Y., Ohm, J.R., Segall, C.A. and Vetro, A., "Standardized extensions of high efficiency video coding (HEVC)," *IEEE Journal of selected topics in Signal Processing,* 7(6), pp.1001-1016, 2013.
39. Nazir, S., Vukobratović, D., Stanković, V., Andonović, I., Nybom, K., and Groenroos, S. "Unequal error protection for data partitioned H. 264/AVC video broadcasting," *Multimedia Tools and Applications,* 74(15), 5787-5809, 2015.
40. Calabuig, J., Monserrat, J.F., Gozàlvez, D. and Gòmez-Barquero, D., "AL-FEC for streaming services in lte e-MBMS," *EURASIP Journal on Wireless Communications and Networking,* 2013(1), p.73.
41. LTE Broadcast - Lessons Learned from Trials and Early Deployments, LTE Alliance, whitepaper, 2016. (http://www.expway.com/lte-broadcast-lessons-learned-from-trials-and-early-deployments/)
42. 3GPP TS 23.002, 14.1.0, "Network architecture."
43. Dahlman, E., Parkvall, S. and Skold, J., "4G: LTE/LTE-advanced for mobile broadband," *Academic press,* 2013.
44. Pu, W., Zou, Z. and Chen, C.W., "Video adaptation proxy for wireless dynamic adaptive streaming over HTTP," *IEEE Packet Video Workshop 2012,* pp. 65-70, 2012.

45. El Essaili, A., Schroeder, D., Steinbach, E., Staehle, D. and Shehada, M., "QoE-based traffic and resource management for adaptive HTTP video delivery in LTE," *IEEE Transactions on Circuits and Systems for Video Technology,* 25(6), pp. 988-1001, 2015.
46. Zhang, W., Wen, Y., Chen, Z. and Khisti, A., "QoE-driven cache management for HTTP adaptive bit rate streaming over wireless networks," *IEEE Transactions on Multimedia,* 15(6), pp.1431-1445, 2013.
47. Erman, J., Gerber, A., Ramadrishnan, K.K., Sen, S. and Spatscheck, O., "Over the top video: the gorilla in cellular networks," *ACM SIGCOMM 2011,* pp. 127-136), 2011.
48. Amram, N., Fu, B., Kunzmann, G., Melia, T., Munaretto, D., Randriamasy, S., Sayadi, B., Widmer, J. and Zorzi, M., "QoE-based transport optimization for video delivery over next generation cellular networks," *IEEE Int'l Symp. Computers and Communications ISCC*, pp. 19-24, 2011.
49. Bouras, C., Kanakis, N., Kokkinos, V. and Papazois, A., "AL-FEC for streaming services over LTE systems," *IEEE Wireless Personal Multimedia Communications (WPMC 2011*, 2011.
50. Capozzi, F., Piro, G., Grieco, L. A., Boggia, G., and Camarda, P. "Downlink packet scheduling in LTE cellular networks: Key design issues and a survey," *IEEE Communications Surveys and Tutorials,* 15(2), 678-700, 2013.
51. Piro, G., Grieco, L. A., Boggia, G., Fortuna, R., Camarda, P., "Two-level downlink scheduling for real-time multimedia services in LTE networks," *IEEE Transactions on Multimedia,* 13(5), 1052-1065, 2011.
52. Lai, W. K., Tang, C. L., "QoS-aware downlink packet scheduling for LTE networks," *Computer Networks*, 57(7), 1689-1698, 2013.
53. Su, G. M., Su, X., Bai, Y., Wang, M., Vasilakos, A. V., and Wang, H. "QoE in video streaming over wireless networks: perspectives and research challenges," *Wireless Networks,* 22(5), 1571-1593, 2016.
54. Araniti, G., Condoluci, M., Militano, L. and Iera, A., "Adaptive resource allocation to multicast services in LTE systems," *IEEE Transactions on Broadcasting,* 59(4), pp.658-664, 2013.
55. Chen, J., Chiang, M., Erman, J., Li, G., Ramakrishnan, K.K. and Sinha, R.K., "Fair and optimal resource allocation for LTE multicast (eMBMS): Group partitioning and dynamics," *IEEE INFOCOM 2015,* pp. 1266-1274, 2015.
56. Lau, C.P., Alabbasi, A. and Shihada, B., "An efficient live TV scheduling system for 4G LTE broadcast," *IEEE Systems Journal,* 11(4), pp.2737-2748, 2017.
57. Bouras, C., Kanakis, N., Kokkinos, V. and Papazois, A., "Embracing RaptorQ FEC in 3GPP multicast services," *Wireless Networks,* 19(5), pp.1023-1035, 2013.
58. IETF RFC 5170, "Low Density Parity Check (LDPC) Staircase and Triangle Forward Error Correction (FEC) Schemes," https://www.rfc-editor.org/rfc/rfc5170.txt
59. Magli, E., Wang, M., Frossard, P. and Markopoulou, A., "Network coding meets multimedia: A review," *IEEE Transactions on Multimedia,* 15(5), pp.1195-1212, 2013.
60. Li, B., Li, H. and Zhang, R., "Adaptive random network coding for multicasting hard-deadline-constrained prioritized data," *IEEE Transactions on Vehicular Technology,* 65(10), pp.8739-8744, 2016.
61. Shin, H. and Park, J.S., "Optimizing random network coding for multimedia content distribution over smartphones," *Multimedia Tools and Applications*, 76(19), pp.19379-19395, 2017.
62. Esmaeilzadeh, M., Sadeghi, P. and Aboutorab, N., "Random linear network coding for wireless layered video broadcast: General design methods for adaptive feedback-free transmission," *IEEE Transactions on Communications,* 65(2), pp.790-805, 2017.
63. Carlucci, G., De Cicco, L. and Mascolo, S., "HTTP over UDP: an Experimental Investigation of QUIC," *ACM Symposium on Applied Computing,* pp. 609-614, 2015.
64. Fragouli, C., Lun, D., Mèdard, M. and Pakzad, P., "On feedback for network coding," *Information Sciences and Systems, CISS 2007,* pp. 248-252, 2007.
65. Sundararajan, J.K., Shah, D., Mèdard, M., Mitzenmacher, M. and Barros, J., "Network coding meets TCP," *IEEE INFOCOM 2009,* pp. 280-288, 2009.
66. Medina Ruiz, H., Kieffer, M., Pesquet-Popescu, B., Medina Ruiz, H., Kieffer, M. and Pesquet-Popescu, B., "TCP and Network Coding: Equilibriu and Dynamic Properties," *IEEE/ACM Transactions on Networking,* 24(4), pp.1935-1947, 2016.
67. Khirallah, C., Vukobratovic, D. and Thompson, J., "Performance analysis and energy efficiency of random network coding in LTE-advanced," *IEEE Transactions on Wireless Communications,* 11(12), pp.4275-4285, 2012.
68. Vukobratovic, D., Khirallah, C., Stankovic, V. and Thompson, J.S., "Random network coding for multimedia delivery services in LTE/LTE-Advanced," *IEEE Transactions on Multimedia,* 16(1), pp.277-282, 2014.

69. Tassi, A. and Khirallah, C. and Vukobratovic, D. and Chiti, F. and Thompson, J. and Fantacci, R., "Resource Allocation Strategies for Network-Coded Video Broadcasting Services over LTE-Advanced," *IEEE Transactions on Vehicular Technology,* 64(5), pp. 2186-2192, May 2015.
70. Tassi, A. and Chatzigeorgiou, I. and Vukobratovic, D., "Resource Allocation Frameworks for Network-coded Layered Multimedia Multicast Services," *IEEE Journal on Selected Areas in Communications,* 33(2), pp. 141-155, Mar. 2015.
71. Tassi, A. and Chatzigeorgiou, I. and Vukobratovic, D. and Jones, A., "Optimized Network-coded Scalable Video Multicasting over eMBMS Networks," *IEEE ICC 2015,* June 2015.
72. Tassi, A. and Khirallah, C. and Vukobratovic, D. and Chiti, F. and Thompson, J. and Fantacci, R., "Reliable rate-optimized video multicasting services over LTE/LTE-A," *IEEE ICC 2013,* June 2013.
73. 3GPP TS 23.501, "System Architecture for the 5G System", www.3gpp.org/DynaReport/23501.htm
74. 3GPP TS 38.300, "NR; Overall Description", www.3gpp.org/DynaReport/38300.htm
75. Ge, C., Wang, N., Foster, G. and Wilson, M., "Toward QoE-Assured 4K Video-on-Demand Delivery Through Mobile Edge Virtualization With Adaptive Prefetching," *IEEE Transactions on Multimedia,* 19(10), pp.2222-2237, 2017.
76. Argyriou, A., Poularakis, K., Iosifidis, G. and Tassiulas, L., "Video Delivery in Dense 5G Cellular Networks," *IEEE Network,* 31(4), pp.28-34, 2017.
77. Golrezaei, N., Molisch, A. F., Dimakis, A. G., and Caire, G. "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Communications Magazine,* 51(4), 142-149, 2013.
78. Poularakis, K., Iosifidis, G., Sourlas, V., and Tassiulas, L., "Exploiting caching and multicast for 5G wireless networks," *IEEE Transactions on Wireless Communications,* 15(4), 2995-3007, 2016.
79. Borcoci, E., Ambarus, T., Bruneau-Queyreix, J., Negru, D. and Batalla, J.M., "Optimization of Multi-server Video Content Streaming in 5G Environment," *Int'l Conf. on Evolving Internet*, 2016.
80. M. Drago, T. Azzino, M. Polese, C. Stefanovic, M. Zorzi, "Reliable Video Streaming over mmWave with Multi Connectivity and Network Coding", *IEEE ICNC 2017,* 2017.
81. Nielsen, J.J., Liu, R. and Popovski, P., "Optimized Interface Diversity for Ultra-Reliable Low Latency Communication (URLLC)," *arXiv preprint*, arXiv:1712.05148.
82. Pervez, F., Adinoyi, A., and Yanikomeroglu, H., "Efficient resource allocation for video streaming for 5G network-to-vehicle communications," *IEEE PIMRC Workshops,* 2017.