Harnessing Cloud and Edge Synergies: Toward an Information Theory of Fog Radio Access Networks

Ravi Tandon and Osvaldo Simeone

The authors consider a hybrid architecture, referred to as Fog-RAN (F-RAN), that harnesses the benefits of, and the synergies between, edge caching and C-RAN. In an F-RAN, edge nodes may be endowed with caching capabilities, while at the same time being controllable from a central cloud processor as in a C-RAN.

ABSTRACT

Edge caching and the centralization of baseband processing by means of the C-RAN architecture are among the most promising and transformative trends in the evolution of wireless networks. A key advantage of C-RAN is the possibility to perform cooperative transmission across multiple edge nodes, such as small cell base stations, thanks to centralized cloud processing. Cloud processing, however, comes at the cost of the potentially large delay entailed by fronthaul transmission between edge and cloud. In contrast, edge caching enables the low-latency transmission of popular multimedia content, but at the cost of constraining the operation of the edge nodes to decentralized transmission strategies with limited interference management capabilities. In order to accommodate the broad range of quality of service requirements of mobile broadband communication, in terms of spectral efficiency and latency, that are envisioned to be within the scope of 5G systems and beyond, this article considers a hybrid architecture, referred to as fog RAN (F-RAN), that harnesses the benefits of, and the synergies between, edge caching and C-RAN. In an F-RAN, edge nodes may be endowed with caching capabilities, while at the same time being controllable from a central cloud processor as in a C-RAN. In this article, an information-theoretic framework is presented that aims to characterize the main trade-offs between performance of an F-RAN, in terms of worst case delivery latency, and its resources: caching and fronthaul capacities.

INTRODUCTION

Edge processing and cloudification are among the most promising trends in the evolution of wireless network architectures toward the specification of fifth generation (5G) systems and beyond. *Edge processing* refers to the placement of storage and computing resources at the network edge, that is, closer to the users. This localization of content and computing caters to low-latency or location-based applications, as well as to multimedia transmission with local content reuse [1]. *Cloudification* amounts to the complementary trend toward the decoupling of physical network elements, such as base stations, from the control and processing logic that is implemented centrally at a cloud processor. The resulting sharing of the control and processing resources of the cloud across multiple network elements yields significant gains in terms of capital and operating expenses, flexibility in ownership models, statistical multiplexing, and interference management [2].

A network architecture based on edge processing is illustrated in Fig. 1a. Here, edge nodes (ENs), such as base stations or eNBs in Long Term Evolution (LTE), are equipped with local caches that can be used to store popular content, most notably multimedia files, with the aim of reducing the delivery latency and the overhead on the backhaul connections to the content server. Edge processing via caching provides an ideal solution for data traffic classes, such as video, characterized by high local content reuse [1]. A scenario with cloudification of the functionalities of the ENs, also known as the cloud radio access network (C-RAN), is depicted in Fig. 1b. In this architecture, the ENs are connected to the cloud processor by so-called *fronthaul* links. Due to the enhanced interference management capabilities afforded by centralized baseband processing at the cloud, which can operate jointly across all connected ENs, C-RANs are particularly well suited to enhance the spectral and cost efficiency of interference-limited dense deployments with less stringent delay constraints [2].

To summarize, while C-RAN provides high spectral efficiencies thanks to cooperative cloud-based transmission, but at potentially large latencies due to fronthaul transmission, edge caching enables the low-latency delivery of popular content, but with limited interference management capabilities because of decentralized baseband processing at the ENs. Recognizing that modern wireless networks, including 5G systems, are expected to cater to a broad range of quality of service requirements for mobile broadband communication, in terms of spectral efficiency and latency, a hybrid architecture was recently advocated [3, 4], which is illustrated in Fig. 2 and referred to as fog RAN (F-RAN). In an F-RAN, ENs may be endowed with caching capabilities, while at the same time being controllable from a central cloud processor. As such, the F-RAN architecture captures the key benefits of central-

Ravi Tandon is with the University of Arizona; Osvaldo Simeone is with the New Jersey Institute of Technology.

0163-6804/16/$25.00 © 2016 IEEE

ized baseband processing and low-latency delivery of C-RAN and edge caching, respectively. It should, however, be noted that the F-RAN architecture does not retain the important C-RAN feature of a reduced deployment cost for the ENs [2], which, unlike for a C-RAN, need to be provided with baseband processing capabilities so as to be able to locally process the cached content.

The main goal of this article is to lay the theoretical foundations for the study of the optimal operation of an F-RAN architecture. Optimal design requires edge caching, fronthaul, and wireless transmission to be jointly designed so as to leverage the discussed synergistic and complementary features of edge processing and virtualization. The resulting design problem is extremely challenging, as it includes the joint optimization of caching, fronthaul, and transmission policies, making a brute force approach prohibitive. To overcome these challenges, in this article, we propose and analyze a *novel information-theoretic framework* with the aims of illuminating the main trade-offs between the system performance in terms of latency on one hand, and the resources available for caching, fronthaul, and wireless transmission on the other, as well as revealing design guidelines for the optimal design of F-RAN via analytical arguments. Examples are offered to illustrate the merits of the proposed approach. References [5, 6] provide the technical details that are omitted here in order to focus on the key ideas.

The rest of the article is organized as follows. We present the proposed information-theoretic model and performance metrics, along with the design space for an F-RAN, which encompasses caching, fronthaul, and transmission policies. We present two case studies that exemplify the analysis afforded by the proposed framework. Generalizations are discussed that concern the impact of imperfect channel state information (CSI) and of the network topology. Finally, we present some concluding remarks and an outlook on open problems and research issues.

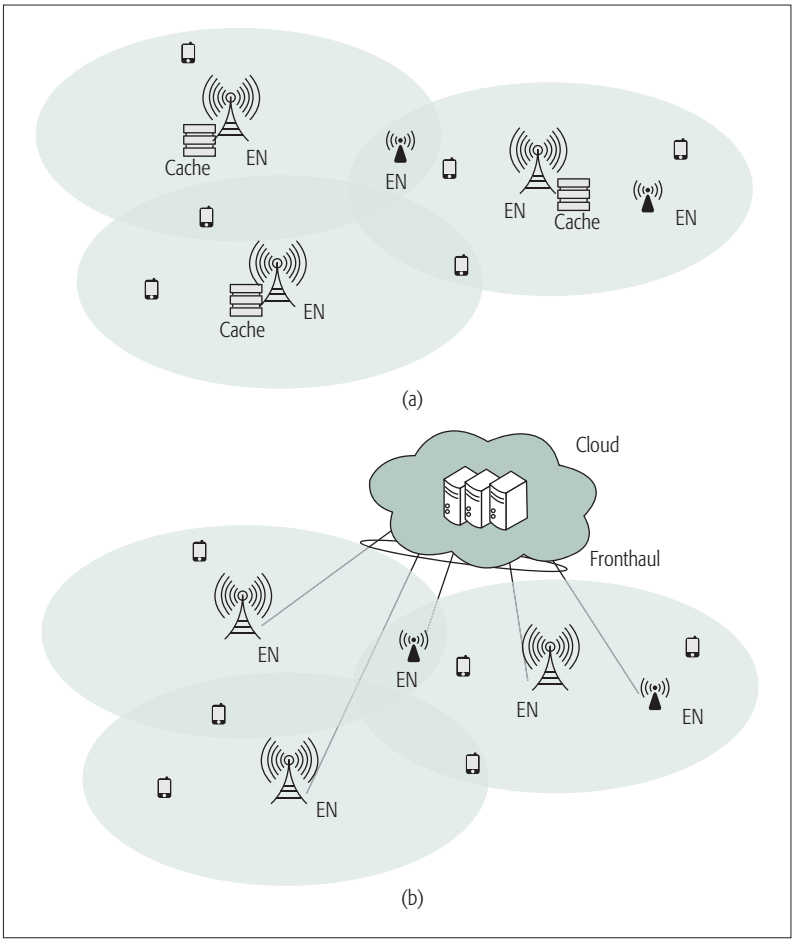

Figure 1. a) A cellular architecture based on edge processing in which some edge nodes are equipped with caches; b) the C-RAN architecture with centralization of edge nodes' baseband processing by means of a cloud processor and a fronthaul network.

Information-Theoretic Model and Design Space

As illustrated in Fig. 3, we consider an F-RAN architecture with M edge nodes (ENs), which can serve a set of K users over a shared wireless channel. The ENs are connected to the cloud by means of fronthaul links of capacity C_F bits per symbol (of the edge wireless channel) for each EN. The capacity C_F is assumed to be fixed, reflecting conventional scenarios in which fronthaul links correspond to dedicated wired connections [2]. Each EN is equipped with a cache of limited size.

We assume the presence of a library of N files, each of length L bits, which represent the content that may be requested by users. As in the majority of related analyses [1, 7, 8], this library of popular files is assumed to remain unchanged during the period of time over which the content of the ENs' caches is not refreshed. For instance, caches may be updated in the early morning, when the traffic load is at a minimum, and kept unmodified for the rest of the day. The period of time in which caches and library are assumed to be fixed encompasses multiple transmission intervals, which are identified by an index $t = 1, 2, \ldots, .$

We focus here on the scenario in which no popularity distribution is available to describe the relative likelihood that one of the files is selected by a user, so all files are equivalent and may potentially be requested [7, 8]. The cache of each EN can store μNL bits for some fractional cache size $0 \le \mu \le 1$.

At each transmission interval t, users issue a vector of requests. We make no assumption on the nature of the time variability of the demands made by users. The collective time-varying wireless CSI $H(t)$ at transmission interval t collects all the channel coefficients that characterize the propagation between all the ENs and the kth user. These coefficients describe the channel profile in the frequency and/or time domain for the given spectral and temporal resources allocated at transmission interval t to a pair of EN and user. For the sake of illustration, following a conventional modeling choice [9], we focus here on the setting in which the channel coefficients are generated independent from an identical continuous distribution. Besides fading, the channel model for the wireless segment includes additive Gaussian noise. Topological constraints are discussed later.

In order to avoid the explicit dependence on the bandwidth of the edge wireless channel, throughout, *rates are measured in bits per symbol of the wireless channel*, and *time metrics are mea-*

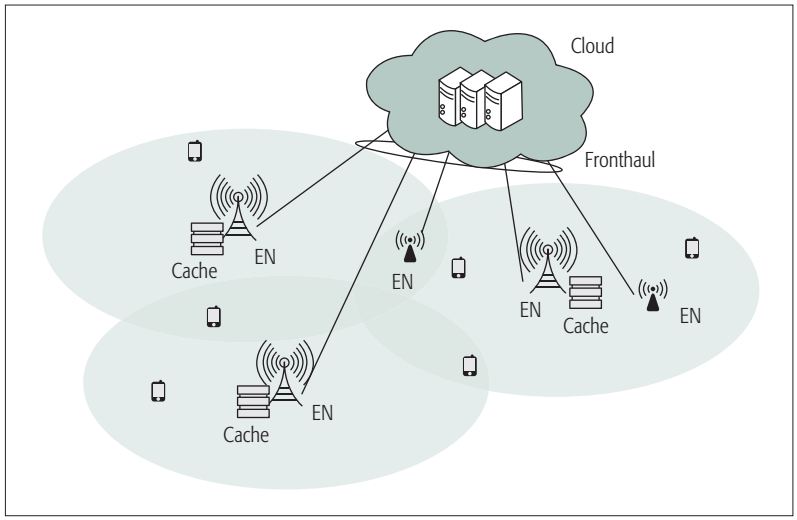

Figure 2. The F-RAN architecture under study that provides a synthesis between edge and cloud processing: ENs may be endowed with caching capabilities as well as with fronthaul connections to the cloud processor.

sured in terms of number of symbols of the wireless channel, or symbols for short.[1]

Design Space. The design problem entails the optimization of the joint caching-fronthaul-transmission policy. Here we summarize the design space under the assumption of full CSI at the ENs and at the cloud. We discuss the impact of imperfect CSI later.

Cache storage policy: The caching policy operates at the discussed timescale over which the set of popular files is expected to remain constant (e.g., one day), which contains many transmission intervals (indexed by t). The caching policy is defined by a function that decides the cache content of each EN. The latter must satisfy the cache capacity constraint, that is, the size of the content stored at each EN cannot exceed μNL bits (Fig. 3). We note that the cache of each EN is populated based solely on the library of files, without knowledge of the instantaneous users' demands as well as without CSI, which vary across transmission intervals t.

A general approach to cache storage is to split each file into a number of fragments of a certain size and to adopt one of the following classes of policies: uncoded caching and coded caching. For *uncoded caching*, each EN stores a subset of the fragments depending on the normalized cache size μ. Uncoded caching policies can create virtual and overlapping clusters of collaborative ENs, where cooperative transmission can be carried out over shared file fragments. For *coded caching*, one can allow for both intra-file coding (i.e., coding within the fragments of a file) and inter-file coding (i.e., coding across different files). Note that coded fragments could also be replicated across ENs to benefit from cooperative transmission as in [10].

Fronthaul policy: The fronthaul policy, as well as the transmission policy, operate separately over each transmission interval t as a function of the instantaneous demands of the users as well as of the CSI of the shared wireless medium. Accordingly, the fronthaul policy is defined by a function of the instantaneous demands and of the CSI that determine the duration T_F of the fronthaul transmission (measured by normalizing with respect to the duration of a symbol of the edge wireless channel). The fronthaul message cannot exceed $T_F C_F$ bits, where C_F denotes the fronthaul capacity as seen above. Note that the fronthaul policy can hence control the fronthaul duration within the given transmission interval.

We can identify two main approaches to the design of the fronthaul policy:

- *Hard-transfer mode*, whereby fragments — coded or uncoded following the classification of caching policies mentioned above — are transferred to the ENs
- *Soft-transfer mode*, whereby the cloud directly encodes the files, producing baseband signals that are quantized and sent over the fronthaul links on the ENs following the C-RAN principle [2]

The design of fronthaul policies in the hard-transfer mode, including the selection of coded or uncoded strategies, follows the guidelines discussed above for caching policies, with the important caveat that the fronthaul policy can adapt the choice of fragments to be sent to ENs to the users' current demands.

To illustrate the design space for soft-transfer mode, consider first the scenario with no caches (i.e., a C-RAN system). Here, the optimization of fronthaul policies is equivalent to that of the transmission over a broadcast channel in which the set of ENs form a multi-antenna transmitter, with the limitation that the encoded baseband signals are subject to the distortion caused by fronthaul quantization. In a more general F-RAN, the design space acquires novel degrees of freedom due to the interplay between coding at the cloud, as in a C-RAN, and coding at the ENs, based on the locally cached content. For instance, each EN may transmit a superposition of the quantized baseband signal received on the fronthaul link and of a function of the cached content. Since the latter is not subject to any quantization noise, this can potentially enhance the performance.

Edge transmission policy: The edge transmission policy, or transmission policy for short, operates on each transmission interval and selects the codewords sent on the wireless channel by all the ENs, and hence also their duration T_E, under an average power constraint given by the parameter P. Note that the codeword transmitted by each EN can depend on the local cache, the received fronthaul message, the instantaneous demands, and the CSI. As elaborated above, the design of transmission policies is strongly interdependent with the caching and fronthaul policies. For instance, with uncoded caching and hard-transfer operation of the fronthaul, the transmission policy design amounts to the problem of coding over a single-hop interference network with arbitrary sets of messages at the ENs.

Latency Metric: Normalized Delivery Time. In order to compare different design choices and to enable system optimization, we adopt the performance metric of the delivery time, that is, the time required by the system to satisfy *arbitrary* users' requests in a given transmission interval. Neglecting the time needed for the cloud and the ENs to register the users' requests, delivery latency is generally affected by the time required for transmission on the two segments of fron-

[1] If the bandwidth of the wireless channel is W, the duration of a symbol is approximately 1/W.

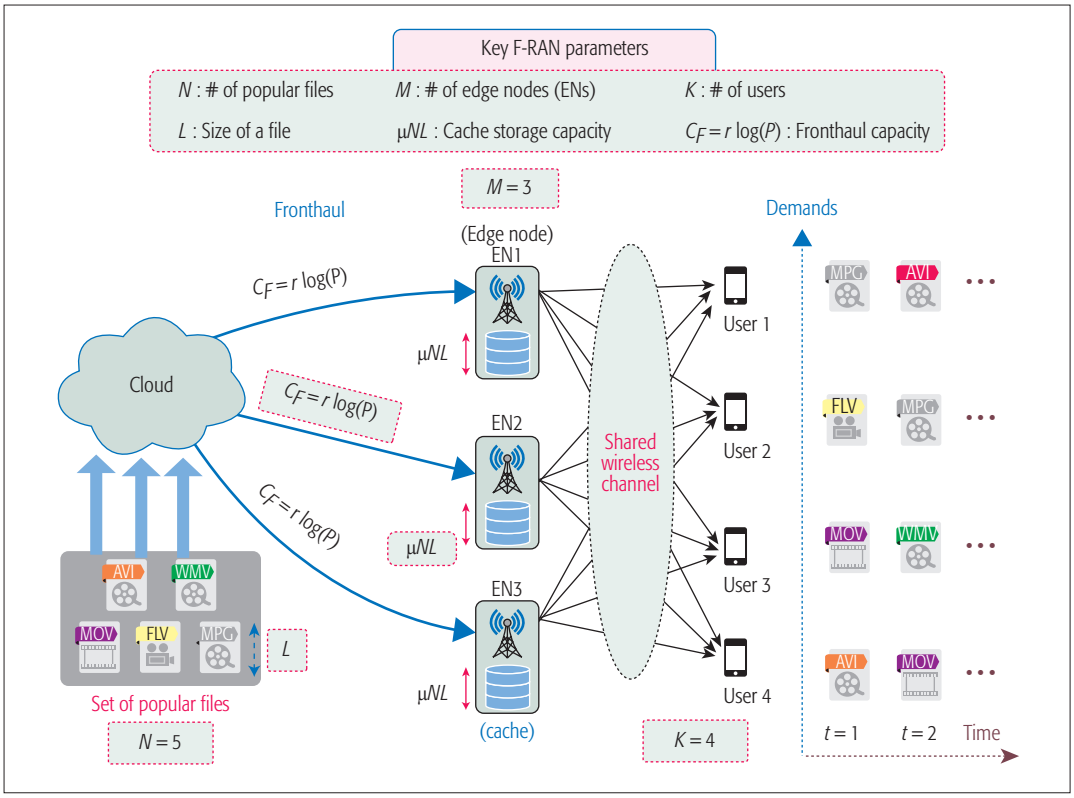

Figure 3. Information-theoretic model for F-RAN.

thaul network and wireless channel. Different assumptions can be made regarding the level of *pipelining* possible between transmissions on the two segments. For example, ENs may immediately start transmitting on the wireless channel while at the same time receiving information on the fronthaul links, which can be causally encoded into the wireless transmission. In this work, we focus on a baseline scenario in which no pipelining is possible, in the sense that fronthaul transmission is followed by wireless transmission.

To elaborate, we first define the delivery time per bit $\Delta(\mu, C_F, P) = (T_E + T_F)/L$, which measures the latency within each transmission interval for the worst case users' request vector, as normalized by the size of the file L. Following the standard Shannon-theoretic framework, the file size L and the blocklengths T_E and T_F are allowed to be arbitrarily large so as to satisfy any desired level of probability of error (see also [11]). The optimal latency performance is in principle obtained by minimizing the delivery time per bit $\Delta(\mu, C_F, P)$ over all possible caching-fronthaul-transmission policies. This optimization is generally prohibitive and is also dependent on all parameters (μ, C_F, P).

With the aim of obtaining analytical insights, we propose a novel tractable metric that retains the key dependence of latency on cache size and fronthaul capacity while adopting a high-signal-to-noise ratio (SNR) approximation in the vein of the by now standard degrees of freedom (DoF) analysis of interference networks [9]. To this end, we let the fronthaul capacity scale with the SNR parameter P as $C_F = r \log(P)$, where r is a parameter that measures the capacity scaling of the fronthaul links' capacity as compared to the wireless channel.

The key idea is to evaluate the relative latency between the F-RAN system under study and that of a *baseline system with no interference and unlimited caching, in which each user can be served by a dedicated EN that has all files*. The delivery time per bit of this ideal system is well known to be $1/\log(P)$, and hence we define the normalized delivery time (NDT) $\delta(\mu, r)$ as the limit of the ratio $\Delta(\mu, C_F, P)/(1/\log(P))$ for large SNR P on the wireless channel. As such, *an NDT of* δ *indicates that the worst case time required to serve any possible request is* δ *times larger than the time that would be needed by the baseline system*. Optimizing over all possible policies yields the minimum NDT $\delta^*(\mu, r)$.

Based on the definitions above, in the proposed framework, the goal of the analysis is the characterization of the novel metric NDT $\delta^*(\mu, r)$ that captures the interplay between latency and resources, that is, the normalized cache storage μ and the fronthaul multiplexing gain r.

Case Studies

Case Study 1: Edge Caching in Interference-Limited Scenarios

We first consider systems with edge caching operating over interference-limited channels with no fronthaul as illustrated in Fig. 1a. We note that the conventional design of cache-aided wireless systems abstracts the contribution of the physical layer by assuming fixed coverage areas and implicitly assumes uncoordinated ENs (see, e.g., [12]). In contrast, it has been recently recognized that, in the presence of shared content in the ENs' caches, the ENs are enabled to use more sophisticated transmission schemes, including coordinated beamforming and precoding. The

Neglecting the time needed for the cloud and the ENs to register the users' requests, delivery latency is generally affected by the time required for transmission on the two segments of fronthaul network and wireless channel. Different assumptions can be made regarding the level of pipelining possible between transmissions on the two segments.

With the aim of obtaining analytical insights, we propose a novel tractable metric that retains the key dependence of latency on cache size and fronthaul capacity while adopting a high-SNR approximation in the vein of the by now standard degrees of freedom (DoF) analysis of interference networks.

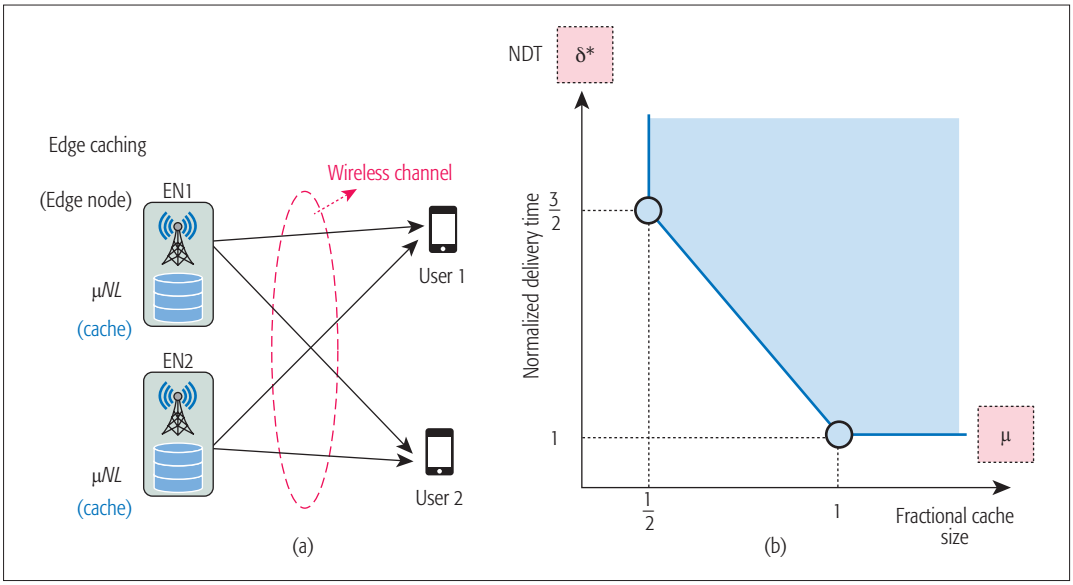

Figure 4. a) Model for an edge caching architecture for $M = 2$ ENs serving $K = 2$ users; b) Trade-off between δ^* (normalized delivery time or NDT) and μ (fractional cache size of the ENs).

interplay between caching and cooperative transmission was first studied in [10], in which it is proposed to store the same erasure-coded packets at all ENs in order to allow for joint beamforming across all ENs. These works are based on dynamic optimization arguments and signal processing. In [8], instead, the cache allocation problem was studied under an information-theoretic framework, from the point of view of DoF analysis, for a scenario with three ENs and users, under the assumption that all the requested files are cached at ENs.

To illustrate the insights afforded by the NDT analysis, we consider the setup in Fig. 3a in which two ENs, labeled EN1 and EN2, are deployed to serve two users. Figure 3b shows the information-theoretically optimal trade-off curve between the NDT δ^* and the fractional cache size μ as obtained in [5] under the constraint of uncoded inter-file caching. Note that this performance trade-off always results in a convex curve [5]. To take some exemplifying operating points on the curve, for $\mu = 1$, both ENs can store all files, and hence full cooperative transmission can take place (i.e., via zero-forcing beamforming for any set of users' requests), yielding $\delta^* = 1$. This implies that the latency performance is the same as that of the mentioned baseline ideal system. On the other hand, at $\mu = 1/2$, which is the smallest cache size to enable delivery of any vector of requests, the NDT increases to $\delta^* = 3/2$ and is achieved via interference alignment [5], revealing the performance loss due to partial caching.

Case Study 2: Fronthaul Processing and Edge Caching for F-RANs

We now elaborate on a full-fledged F-RAN scenario with cloud processing and edge caching for F-RANs as illustrated in Fig. 2. As a case study, we consider the F-RAN topology shown in Fig. 4a, in which the edge nodes EN1 and EN2 are endowed with caches, as discussed in the previous example, but are also connected to the cloud by means of fronthaul links with given capacities. From a signal processing viewpoint, the joint design of beamforming and fronthaul processing in hard-transfer mode, where the latter determines which ENs receive each non-cached file on the fronthaul links, is studied in [13, 14] for a fixed pre-defined cache allocation.

Figures 4b and 4c show the optimal NDT trade-off derived in [6], again under the assumption of uncoded caching. We first note that NDT trade-off identifies two distinct regimes in terms of the fronthaul capacity, a *low-fronthaul capacity regime* with $r \le 1$ and a *high-fronthaul capacity regime* with $r > 1$. In the latter case, the use of both fronthaul and caching resources is necessary in order to obtain the optimal NDT performance, while in the former, if the cache capacity is sufficiently large (i.e., if $\mu \ge 1/2$), it is sufficient to leverage the cache storage resources to achieve the optimal performance.

To provide additional insights on the calculation and significance of the NDT metric, we now briefly discuss the scheme that achieves the NDT $\delta^*(\mu = 0, r) = 1 + 1/r$. The case $\mu = 0$ corresponds to the setting in which the ENs have no cache storage capability. A finite NDT can hence only be achieved by using the fronthaul resources. As discussed, the fronthaul links can be utilized in either hard- or soft-transfer mode. With hard transfer, the cloud can transmit both requested files to each EN, and then the ENs can use the same fully cooperative zero-forcing approach adopted for $\mu = 1$, as discussed above. Since the fronthaul links have capacities $C_F = r \log(P)$ each and $2L$ bits need to be sent to both ENs, the achievable NDT can be computed as $\delta = 1+2/r$. However, hard transfer turns out to be suboptimal in this scenario. The optimal NDT is in fact achieved through a soft-transfer scheme, whereby the cloud implements zero-forcing beamforming and quantizes the resulting baseband signals. It can be shown [6] that this scheme entails a fronthaul latency that equals the edge latency multiplied by $1/r$, since the scheme uses a resolution of around $\log(P)$ bits per downlink sample. As a result, it yields the optimal NDT $\delta^* = 1 + 1/r$.

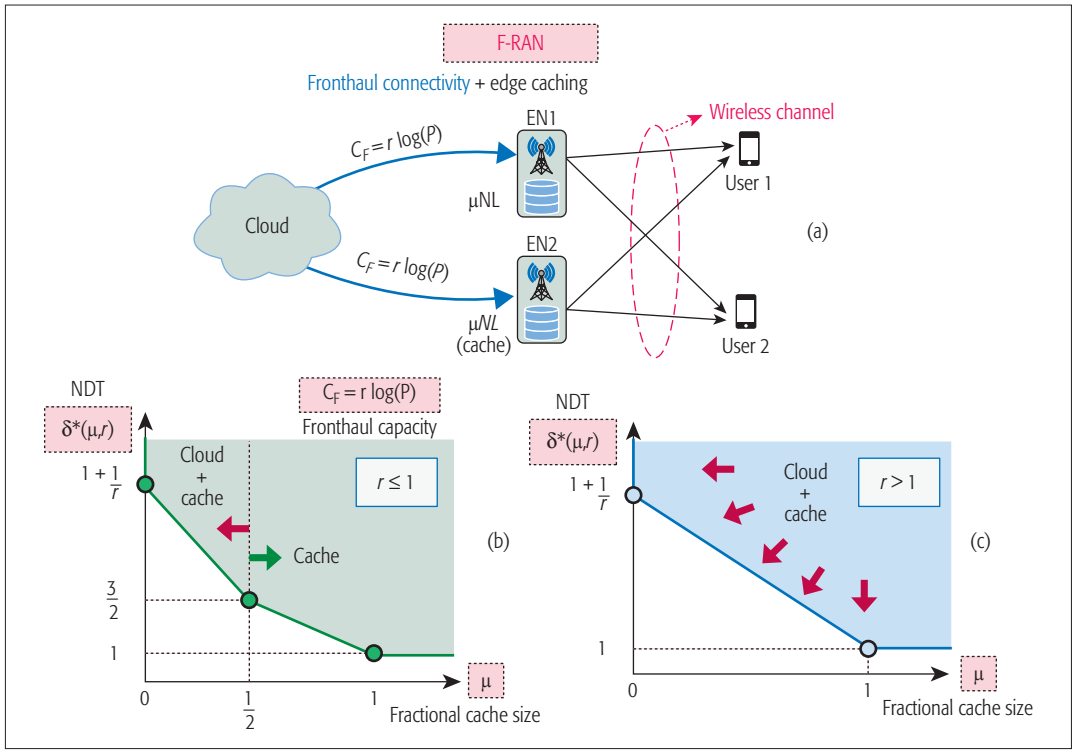

Figure 5. a) Model for an F-RAN system for $M = 2$ ENs serving $K = 2$ users; b–c) Optimal NDT trade-off as a function of μ (fractional cache size per EN) and fronthaul capacity $C_F = r \log(P)$. The trade-off has distinct regimes of operations for two cases: (b) $r \le 1$; (c) $r > 1$.

The CSI acquired at the EN is typically local in the sense that it only pertains the channels describing propagation from the given EN, and not from the other ENs, to the users. As a result, in an F-RAN, the CSI model has the following two unique features: heterogeneous CSI timeliness and global vs. local CSI.

Generalizations

In this section, we discuss two generalizations of the information-theoretic framework studied so far with the aim of accounting for imperfect CSI and for the impact of topology. Other generalizations of interest, not further discussed here, include the investigation of the impact of pipelining of fronthaul and wireless transmissions; the consideration of online caching strategies in which the caches can be updated based on the signals received from the cloud on the fronthaul [15]; and the study of the impact of limited reliability transmission in the finite blocklength regime.

The Impact of Imperfect CSI

In an F-RAN, for both time-division duplex (TDD) and frequency-division duplex (FDD) operations of the wireless channel, CSI is first estimated at the ENs, either directly through uplink training for TDD or indirectly via feedback for FDD, and then conveyed to the cloud through the fronthaul links. Furthermore, the CSI acquired at the EN is typically *local* in the sense that it only pertains to the channels describing propagation from the given EN, and not from the other ENs, to the users. As a result, in an F-RAN, the CSI model has the following two unique features:

- *Heterogeneous CSI timeliness*: CSI acquired at the ENs is more current than the CSI available at the cloud due to the delay in fronthaul transfer between ENs and cloud.
- *Global vs. local CSI*: The ENs have local, more timely, CSI, and the cloud has global, but more delayed, CSI.

We next describe the proposed system model that aims at capturing these aspects, as well as the significant novel challenges that arise from the study of F-RANs with imperfect CSI in terms of both policy design and converse arguments.

As an exemplifying illustration of the main new challenges that arise in the design of caching, fronthaul, and transmission policies due to the heterogeneity of the CSI timeliness at ENs and cloud, we now consider the $M = 2$-EN and $K = 2$-user example studied above in which possibly delayed CSI is available at the ENs. We focus here for simplicity on the setup with $r = 0$ so that only caching, and no cloud transmission, is considered. The resulting NDT trade-off is shown in Fig. 6. The figure illustrates the impact of increasing delays at the ENs on the NDT, starting from no delay, to delay as large as coherence time or "stale" CSI, ending with no CSI at the ENs. Focusing on the operating point with $\mu = 1/2$, we observe that when CSI is timely, the minimum NDT of 3/2 is achieved via a scheme based on interference alignment as discussed above and in [12]. Moreover, when CSI is outdated, the effect of "stale" CSI is reflected in an increase of the NDT to 5/3, which can be achieved via a transmission scheme that uses "stale" CSI [5]. Finally, when there is no CSI, an NDT of 2 is achieved by independent transmissions to each of the users (e.g., using time division), requiring twice the time compared to full CSI.

Impact of Network Topology

Here, we focus on *a general network topology*, in which a user may be in the coverage of only a subset of ENs, hence receiving at negligible power for the rest of the ENs. This scenario captures the operation of larger-scale networks in which ENs cover different, but possibly overlapping, areas. Specifically, the system model discussed above is modified here by allowing the channel gains

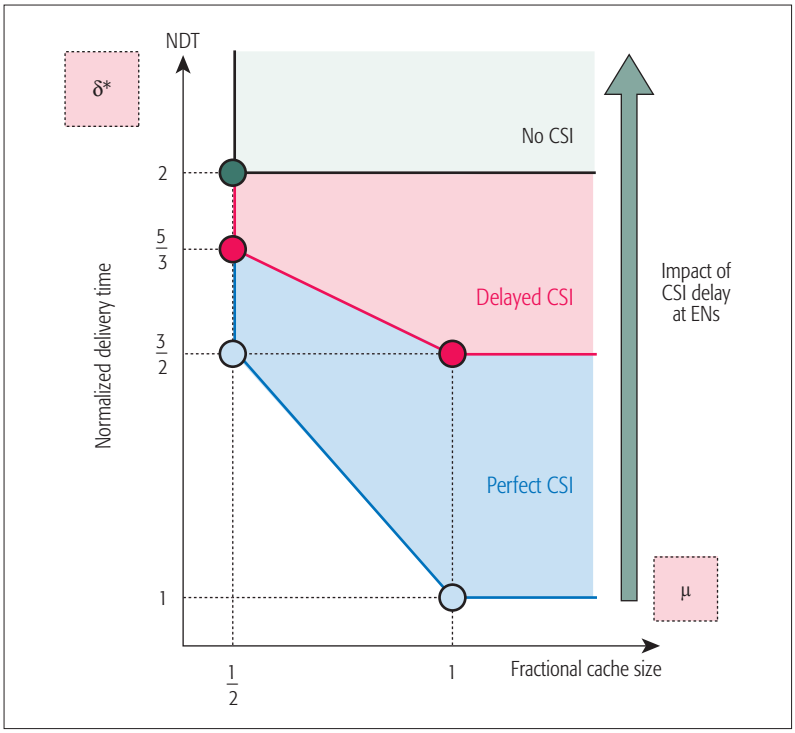

Figure 6. Impact of CSI on latency for F-RANs.

between given pairs of ENs and users to be zero, signifying the fact that a user is outside the coverage of an EN. We characterize the connectivity of a given topology by the parameter ℓ, which denotes the minimal number of ENs that cover any user u. In order to avoid uninteresting pathological cases and maintain tractability, we further assume that $M = K$ and each EN covers the same number of users. For instance, the choice $\ell = M$ corresponds to a fully connected network and $\ell = 1$ corresponds to non-interfering point-to-point channels.

We provide now an example that shows that coded caching can provide unbounded gains in general topologies. Consider a ring topology with $\ell = 2$, in which there are three equally spaced ENs, and three users placed between two successive ENs and connected only to the two nearby ENs. Assume that $\mu = 1/2$ and that $r = 0$. With uncoded caching, each EN can hence at most store at most half of every file. Therefore, it can be seen that due to the limited connectivity, there are users' requests that cannot be met, yielding an unbounded NDT. For instance, if EN1 and EN2 store the first half of a given file and EN3 the other, a user connected only to EN1 and EN2 cannot recover the file if requested. Instead, with coded caching, we split a file A into two equal size fragments A_1 and A_2. EN1 can cache the first half A_1 of the file, EN2 can cache the second half A_2, and EN3 the intra-file coded fragment $A_1 \oplus A_2$ given by the XOR of the two fragments. With this coded caching scheme, the NDT is finite since a user attached to *any two* ENs can recover any file. Note that this amounts to the use of an $(n, k) = (3, 2)$ MDS code.

Concluding Remarks and Outlook

The F-RAN architecture leverages the synergies between cloud processing and edge caching to offer performance advantages in terms of latency and spectral efficiency. In this article, we have introduced an information-theoretic framework that aims at capturing the key trade-off between delivery latency and main system resources, namely fronthaul capacity and caching storage capacity. We have provided a number of use cases, exemplifying examples, and open problems. In presenting the framework at a high level, the authors hope to stimulate research on the topic.

References

[1] E. Bastug, M. Bennis, and M. Debbah, "Living on the Edge: The Role of Proactive Caching In 5G Wireless Networks," *IEEE Commun. Mag.*, vol. 52, no. 8, 2014, pp. 82–89.

[2] A. Checko *et al.*, "Cloud RAN for Mobile Networks: A Technology Overview," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 1, 2014, pp. 405–26.

[3] Q. Li, H. Niu, A. Papathanassiou, and G. Wu, "Edge Cloud and Underlay Networks: Empowering 5G Cell-Less Wireless Architecture," *Proc. European Wireless 2014*, 2014, pp. 1–6.

[4] S.-H. Park, O. Simeone, and S. S. (Shitz), "Joint Optimization of Cloud and Edge Processing for Fog Radio Access Networks," *Proc. IEEE Int'l. Symp. Info. Theory*, 2016.

[5] A. Sengupta, R. Tandon, and O. Simeone, "Cache Aided Wireless Networks: Tradeoffs between Storage and Latency," *Proc. Conf. Info. Sciences and Systems*, 2016.

[6] R. Tandon and O. Simeone, "Cloud-Aided Wireless Networks with Edge Caching: Fundamental Latency Trade-Offs in Fog Radio Access Networks," *Proc. IEEE Int'l. Symp. Info. Theory*, 2016.

[7] M. A. Maddah-Ali and U. Niesen, "Fundamental Limits of Caching," *IEEE Trans. Info. Theory*, vol. 60, no. 5, 2014, pp. 2856–67.

[8] –, "Cache-Aided Interference Channels," *Proc. IEEE Int'l. Symp. Info. Theory*, June 2015, pp. 809–13.

[9] S. A. Jafar, "Interference Alignment – A New Look at Signal Dimensions in a Communication Network," *Foundations and Trends in Commun. and Info. Theory*, vol. 7, no. 1, 2010, pp. 1–34.

[10] A. Liu and V. K. Lau, "Exploiting Base Station Caching in MIMO Cellular Networks: Opportunistic Cooperation for Video Streaming," *IEEE Trans. Signal Processing*, vol. 63, no. 1, pp. 57–69, 2015.

[11] Y. Liu and E. Erkip, "Completion Time in Multi-Access Channel: An Information Theoretic Perspective," *Proc. IEEE Info. Theory Wksp.*, 2011, pp. 708–12.

[12] K. Shanmugam *et al.*, "Femtocaching: Wireless Content Delivery Through Distributed Caching Helpers," *IEEE Trans. Info. Theory*, vol. 59, no. 12, 2013, pp. 8402–13.

[13] X. Peng *et al.*, "Joint Data Assignment and Beamforming for Backhaul Limited Caching Networks," *Proc. IEEE PIMRC*, 2014, pp. 1370–74.

[14] M. Tao *et al.*, "Content-Centric Sparse Multicast Beamforming for Cache-Enabled Cloud RAN," to appear, *IEEE Trans. Wireless Commun.*

[15] R. Pedarsani, M. A. Maddah-Ali, and U. Niesen, "Online Coded Caching," *Proc. IEEE ICC*, 2014, pp. 1878–83.

Biographies

Ravi Tandon (tandonr@email.arizona.edu) received his B.Tech. degree in electrical engineering from the Indian Institute of Technology, Kanpur in 2004 and his Ph.D. degree in electrical and computer engineering from the University of Maryland, College Park (UMCP) in 2010. From 2010 to 2012, he was a postdoctoral research associate in the Department of Electrical Engineering at Princeton University. He is currently an assistant professor in the Department of Electrical and Computer Engineering at the University of Arizona. Prior to joining the University of Arizona in fall 2015, he was a research assistant professor at Virginia Tech with positions in the Bradley Department of ECE, the Hume Center for National Security and Technology, and the Discovery Analytics Center in the Department of Computer Science. He was a co-recipient of the Best Paper Award at IEEE GLOBECOM 2011. He was nominated for the Graduate School Best Dissertation Award, and also for the ECE Distinguished Dissertation Fellowship Award at UMCP. His current research interests include information theory and its applications to wireless networks, communications, security and privacy, distributed storage systems, machine learning, and data mining.

Osvaldo Simeone [F] (osvaldo.simeone@njit.edu) received his M.Sc. degree (with honors) and Ph.D. degree in information engineering from Politecnico di Milano, Italy, in 2001 and 2005, respectively. He is currently with the Center for Wireless Information Processing (CWiP), New Jersey Institute of Technology (NJIT), Newark, where he is a professor. His research interests concern wireless communications, information theory, optimization, and machine learning. He was a co-recipient of the 2015 IEEE Communication Society Best Tutorial Paper Award, and Best Paper Awards at IEEE SPAWC 2007 and IEEE WRECOM 2007. He currently serves as an Editor for *IEEE Transactions on Information Theory*.