Video Delivery in Dense 5G Cellular Networks

Antonios Argyriou[†], Konstantinos Poularakis[*], George Iosifidis[‡], Leandros Tassiulas[*]
[†]Department of Electrical and Computer Engineering, University of Thessaly, Greece
[*]Department of Electrical Engineering, and YINS, Yale University, USA
[‡]School of Computer Science and Statistics, Trinity College Dublin, and CONNECT Centre, Ireland

***Abstract*—One of the main challenges for mobile network operators today is the efficient delivery of mobile video-on-demand (VoD) services. In order to satisfy their requirements, we must revolutionize the way we manage the densely deployed resources in 5G wireless networks, and tailor their operation for video content delivery. Two key approaches that leverage the density of the 5G infrastructure are: (i) the systematic usage of in-network storage resources available in base stations (BSs) for the deployment of smart caching policies, and (ii) the video delivery optimization techniques which leverage the novel multi-tier, dense, and heterogeneous structure of 5G systems. We first provide an overview of the main features of the densely deployed 5G networks that are expected to shape mobile VoD services. Next, we present specific solutions that have been recently proposed for each one of the approaches above, and discuss the main lessons learned. Finally, we discuss a number of open problems for video delivery in dense 5G systems.**

I. Introduction

5G is the next generation of wireless systems that aspires to become a paradigm shift and not just an incremental version of existing cellular networks. One of its key features will be the extremely dense deployment of base stations of various form factors. This heterogeneous and multi-tier architecture is envisaged to deliver orders of magnitude higher throughput, lower delay, and energy efficiency. This approach will enable 5G systems to support a significantly enlarged and diversified bouquet of applications. Clearly, one of the main tasks of these ultra-dense 5G networks will be to satisfy mobile video demand ranging from HD to UHD and novel multimedia services, which is driving the mobile data traffic growth.

The content consumption habits of mobile users are rapidly changing. Nowadays, users spend more time on their mobile devices than their TVs. To exemplify, according to Nielsen ratings, 84% of mobile devices are used while watching TV programs offered by Video on Demand (VoD) platforms such as Netflix, Hulu, and Apple TV. This is facilitated by the latest developments in video delivery technology, ranging from adaptive streaming mechanisms to sophisticated video codecs, that aim to revolutionize the user viewing experience. These developments place huge pressure to Mobile Network Operators (MNOs) which need to leverage these novel technologies and develop clean-cut video delivery solutions to satisfy the unprecedented, in volume and performance, requests of users.

The most popular types of requests are video streaming and video content downloading. In both cases, the user experience is enhanced through the delivery of high resolution video content that is delivered with low delay, and rendered smoothly and without interruptions (playback stalls), a condition that is more challenging to satisfy in the case of streaming. 5G operators are planning to satisfy these requests through the dense deployment of wireless infrastructure that can both increase the capacity but also the energy efficiency of video delivery. At the same time however, they must suppress their operating expenditures. This can be achieved by utilizing efficiently the available physical resources and by reducing the energy consumption of the deployed active network components. Oftentimes, the above objectives are contradicting, and a proper balance must be carefully selected, a task that becomes more intricate the denser the networks become.

To that end, two novel features of ultra-dense 5G systems are expected to play a key role in efficient video delivery:

- the in-network storage that is available today in abundance and at low-cost, and
- their multi-tier, increasingly heterogeneous and dense architecture of the Radio Access Network (RAN).

Storage, a hitherto underutilized resource in wireless networks, can be deployed at large in 5G systems. This is essential for the transformation of cellular networks to mobile video delivery platforms: storage can substitute expensive capacity investments and, at the same time, improve the user experience through proper caching policies. The dense multi-tier heterogeneous structure, on the other hand, enables concurrent and multi-path connections from the core network to the end users. This creates novel opportunities for optimizing the delivery of video content, either by using enhanced physical layer techniques or smart routing approaches.

In this paper we first provide a brief overview of the key features of the dense 5G system that are expected to revolutionize video delivery, and of the latest developments in mobile VoD technology (Section II). Then, we analyze how the growing demands of users can be satisfied through two classes of solutions: (i) advanced network-aware caching strategies for *video delivery through downloading or streaming* (Section III), and (ii) holistic end-to-end optimization for *video delivery through streaming* that features proactive and dynamic channel-aware algorithms (Section IV). Finally, Section V summarizes and presents key future research questions.

II. Background: 5G and Mobile Video

A. Dense 5G System Architecture

5G networks are expected to be complex due to coexistence of diverse wireless communication technologies and the ultra-dense deployment of various types of base stations and other

This work was supported in part by a research grant from the Science Foundation Ireland (SFI) under Grant Number 13/RC/207, and the National Science Foundation under Grant CNS 1527090. K. Poularakis acknowledges the Bodossaki Foundation, Greece, for a postdoctoral fellowship.

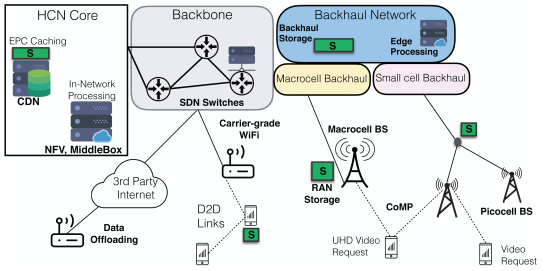


Fig. 1. Overview of ultra-dense 5G networks. Caching of video content (denoted "S") will be realized at the network core, at the RAN, base stations (BSs) of different form-factors and even at mobile devices. Advanced physical layer techniques, such as CoMP and D2D transmissions, will proliferate and support Ultra High Definition (UHD) video delivery. SDN and NFV capabilities will allow for per flow routing and bandwidth provisioning, and sharing of network and storage resources.

radio elements. In the sequel we describe the 5G model placing emphasis on the aspects pertinent to the density of wireless network infrastructure and video delivery. An overview of the envisaged system architecture is presented in Fig. 1.

Multi-tier Architectures and Storage. 5G networks will be characterized by the high density of base stations (BS) [1]. Various types of small cell base stations (SCBS), will underlay the typical macro base stations (MBS) and give rise to the multilayer heterogeneous cellular network (HCN). Therefore, BS density means that users will concurrently be within range of multiple BSs, often of different types, and there will exist multiple paths connecting them to the requested content. Each path will yield a possibly different cost for the operator, and different performance for the users. For example, SCBSs will often be connected to the network core through long-range and low-capacity backhaul links, but typically will serve few users. Hence, delivering video through a SCBS can increase the actual throughput but it might also introduce high latency. On the other hand, this multilayer dense structure increases intra-cell interference which, in turn, reduces the spectral efficiency of wireless transmissions within the cells, especially for SCBSs located close to the MBS. This often results in highly fluctuating levels of interference, and therefore complicates the resource allocation decisions that rely on such time-sensitive information.

Another novel feature of 5G systems is the utilization of in-network storage. This has become a readily available and low-cost commodity, and as such, can be leveraged to improve the performance of 5G systems. For example, popular content caching at the network core can reduce the expensive off-network bandwidth consumption, and improve the end-user experience. Similarly, caching at the RAN or even at the base stations, expedites content delivery and alleviates the problem of limited capacity of the SCBS backhaul links [2]. Besides, the large storage capacity of modern user devices allows them to cache content and distribute it through device-to-device (D2Ds) communications. Operators have begun experimenting with such solutions, following similar developments in wireline networks.

Network Sharing. Finally, a pivotal idea in 5G systems is the extensive employment of cooperation mechanisms, in all possible levels. From the MNOs' perspective, the hardware abstraction and virtualization offered by software-defined networking (SDN) and network function virtualization (NFV) offer new network sharing solutions that can suppress their expenditures especially in urban areas of dense deployments. For example, operators can share their storage resources to cache video content, or share the cached items and even co-design their caching policies [3]. From the end users' perspective, the proliferation of advanced handheld devices facilitates cooperative solutions such as direct exchange of video content and relaying of video streams. Such D2D solutions, either in fully autonomous mode or in cooperation with the operators, are expected to play a key role in 5G networks [4]. Mobile video delivery in particular, can greatly gain by such schemes due to the inherent locality of video content popularity.

B. Mobile Video Technology

In this new era, it is crucial for 5G operators to understand the latest advances in video technology, as well the factors that shape user satisfaction (utility) from mobile video viewing. Both of these aspects need to be taken into account when designing their video delivery mechanisms that may be based either on downloading or the more sophisticated *adaptive streaming* techniques. Video downloading performance depends on the throughput of the communication link that affects eventually the video startup delay. On the other hand adaptive streaming has a more complicated set of performance metrics as we discuss next.

Adaptive Streaming. One class of mobile video delivery techniques that are currently gaining increasing popularity are the adaptive streaming protocols. Prominent examples include dynamic adaptive streaming over HTTP (DASH) [5], and HTTP Live Streaming (HLS). The main idea is that a video file is stored as a sequence of smaller segments with a typical duration of a few seconds. Each segment is available at different quality in terms of video signal SNR, spatial resolution, frame rate, or any combination of these different quality metrics. Based on the actual end-to-end throughput that each user achieved during the delivery of the current segment, the protocol determines the quality (and hence the size) of the next segment. The goal is to maximize the video quality while minimizing the two most important elements of the video streaming Quality of Experience (QoE): (i) the number of video *playback stalls*, and (ii) the *playback stall duration*. However, adaptive streaming protocols are not aware of the network-side state and objectives. For example, they can induce high energy consumption by overloading certain network components. It remains an open question to design proactive streaming services that ensure high user QoE and meet the operators' goals.

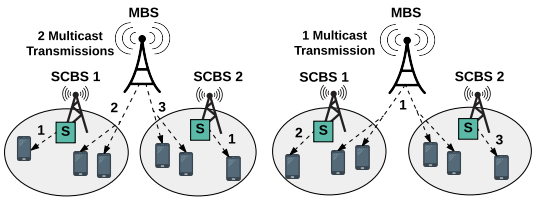
Scalable Video Coding. Unlike data, video can be available in several different quality levels described in the last paragraph. Content providers, traditionally make available multiple video versions encoded at various rates. These versions offer different user experience and have different storage and network bandwidth requirements. Alternatively, layered video encoding [6] (or scalable video coding, SVC) can offer the

same flexibility with lower storage requirements regardless of the delivery mechanism (downloading or streaming). With SVC each video is encoded into different layers which, when combined, produce a quality that increases as more layers are used (e.g. higher frame rate, higher resolution). This technique introduces an encoding overhead (hence, requires more bandwidth) but offers network flexibility since the layers of each file can be cached at the available storage reservoirs, Fig. 1, and routed over different paths. From the operator's point of view, it is important to determine which technique (versions or layers) is more suitable for satisfying the diverse user requirements. This problem that has been investigated for wireline networks [7], takes an entirely new twist in the multi-layer heterogeneous 5G systems [8].

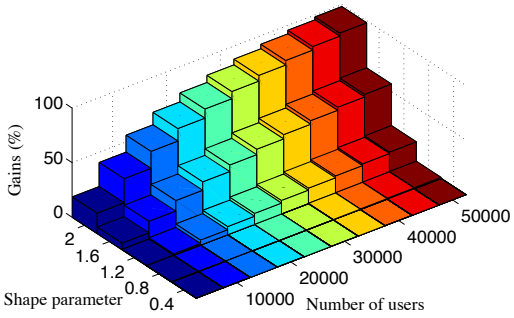
Video Codecs. In addition to the explosive growth in on-demand video, we are in the middle of a transition from high definition video to ultra high definition (UHD) video, with higher 4K/8K resolution, higher 10-bit color accuracy, and higher dynamic range. HEVC/H.265 was built to meet the previous requirements and match the capabilities of future screens. HEVC/H.265 is nearly 50% more efficient than H.264 in terms of bitrate for the same video quality at the cost of higher complexity at the encoder and decoder. Beyond HEVC/H.265, there are other significant developments in video codecs that depart from the past. AV1 will be the first codec released by the Alliance for Open Media (AOM). It is designed to replace Google VP9 and to compete with HEVC/H.265. One of its competitive advantages is the Alliance membership, which ensures the royalty-free deployment of AV1 playback in browsers, mobile devices, and smart TVs, as well as the distribution of AV1-encoded content by YouTube, Netflix, and Amazon. Other video coding standards based on HEVC/H.265 and AV1 may also proliferate in the future. These include multiview video coding like the MV-HEVC extension, and 3D video extensions of H.264 and HEVC/H.265, technologies that use multiple independent video streams. It is critical for the operators to be fully aware of the video content that they deliver so that they can configure and optimize their networks accordingly. For example, during streaming, the smaller video files of the newer codecs are more robust to network performance fluctuations, while the same video content consumes fewer resources and induces lower costs. In the sequel, we discuss in detail such solutions.

III. Network-aware Video Caching

Proactive in-network caching has been traditionally considered a very efficient method for optimizing content delivery. It enhances the user experience (reduces delays), and lowers the network costs by decreasing the bandwidth consumption and the network energy expenditures. Given the vast number of video files, the challenge here is to design the *video caching policy*, i.e., decide where to cache each video file and what encoding rate (quality) to use for it. For this caching problem, we consider video delivery through *downloading*. Hence, the video performance metric of interest is the video startup delay. This is a crucial QoE metric regardless of the delivery method (streaming or downloading) and correlates positively to the



(a) Multicast and caching are coupled.



(b) Impact of diversity of user demand.

Fig. 2. **(a)**: Left: In typical caching, both SCBSs cache video 1 (most popular) and two multicast transmissions are required for the remaining video requests. Right: In multicast-aware caching, video 1 requests are transmitted with one multicast, and local requests are satisfied by low-cost SCBS caches. **(b)**: Energy benefits of multicast-aware caching, over conventional caching, increase when user demand becomes massive and video popularity distribution more steep (captured by the shape parameter of a Zipf probability distribution [11]).

engagement of users. We briefly touch upon *video streaming* in this section by proving new and unpublished results, while we discuss in detail methods for optimizing it through an end-to-end network optimization in Section IV.

Even when we consider video downloading, caching is already a computationally hard problem and various solutions have been previously proposed for wireline networks. However, video caching in dense 5G HCNs is substantially different because of: (i) the heterogeneous and dense multi-tier network structure, (ii) the video demand which is often expected to be massive (e.g., during sport events), and (iii) the user requests which, unlike other cases, have significant elasticity in terms of video quality. The mobile video delivery solutions discussed in the section fit perfectly into the SDN paradigm. SDN allows the operators to employ extremely fine-grained traffic engineering control. For example, routing and bandwidth allocation can be designed on a per-flow basis instead of using generic origin-destination criteria. Therefore, for each video service that a user receives, the network can employ a different video delivery policy.

A. How to Design Video Caching Policies in HCNs?

The first question we investigate in this context is how to optimally design a policy for caching video content at the network edge, i.e., at various small cell BSs. This idea, originally proposed in [2], is very promising [9] as it reduces the required backhaul capacity, a huge cost factor for the deployment of SCBS. Unlike prior works, we studied this problem by taking into account the limited wireless capacity of SCBS. Clearly, caching a file at a base station that does

not have enough capacity to deliver it, is of no practical use. This coupling is more pronounced when user demand is large enough and/or the BS are of very small form factor, e.g., a pico or femto eNB, as is the case in ultra-dense 5G systems.

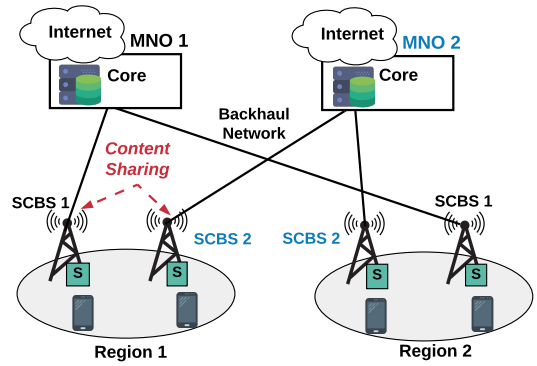
In particular, in [10] we analyzed this scenario by employing facility location theory, where each facility represents a cached item at an SCBS. Our goal was to find the servicing policy that maximizes the SCBS cache hit ratio and therefore reduces the number of energy-consuming MBS transmissions. The obtained solution yields not only the caching but also the routing decisions. In other words, it dictates where to cache each item and how to route each request. The model considers important features of the system, such as the storage and bandwidth heterogeneous constraints of the base stations, and the spatial variations in content popularity. We reduced this problem to the Unsplittable Hard-Capacitated Metric Facility Location instance and, using this mapping, we designed new approximation algorithms. Through extensive evaluation it was found that these joint routing and caching algorithms perform very well (10% optimality gap in practice), and achieve an MBS load reduction that outperforms by almost 40% the caching policies that are link-capacity unaware.

A similar joint optimization approach was followed in [11] where we co-designed caching and multicast policies. Multicast is a particularly efficient technique as it exploits the broadcast nature of the wireless medium and replaces multiple unicast transmissions. In 5G systems, where the mobile data demand will often be massive, and an increasing number of multimedia services (e.g., social networking platforms) will employ the one-to-many communication paradigm, multicast is expected to have a special role. Clearly, this will affect the policies for caching video files at the base stations or other locations at the RAN, as it is explained with the example of Fig. 2(a). Therefore, we need to understand how is caching affected by multicast, and how can we combine these two functions to reduce the network's energy consumption.

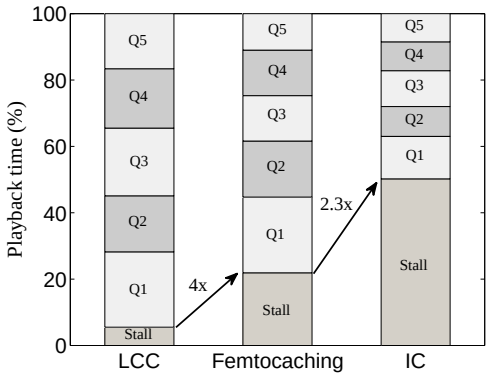
To address this question, we introduced a discrete optimization problem that jointly devises the caching policy (where to cache each video file) and the multicast strategy (which file and from which BS to transmit). We showed that this multicast-aware caching problem is NP-Hard even to approximate within a factor of $O(\sqrt{N})$, where N is the number of SCBS per macrocell. We used randomized-rounding techniques and developed solutions with performance guarantees under the assumption of (bounded) capacity expansion. We also proposed heuristic solutions (requiring no expansion) that perform very well in practice. Trace-driven evaluation showed that these joint designs achieve energy savings that largely exceed the conventional multicast-unaware caching policies. Moreover, these benefits increase as the demand becomes more heavy (number of requests/sec) and more steep (a few video files attract most of the demand), Fig. 2(b).

B. How to Cope with the Elasticity of Video Quality Demand?

A second question is how to better satisfy the user requests, given that these are often elastic in terms of video encoding quality. Delivering high video quality to the users



(a) Illustrating cooperative caching across two MNOs.



(b) Impact of Cooperative Caching on the duration of playback stalls

Fig. 3. **(a)**: Cooperative caching architecture where local caches at nearby SCBSs are jointly used to serve the area's demand. **(b)** impact of cooperative caching on video streaming performance. The proposed layer-aware cooperative caching (LCC) strategy reduces the stalls and increases the quality of streamed videos (Q5>Q4>...>Q1) compared to the indepenent caching strategies (IC) or other cooperative solutions (e.g., Femtocaching [2]) which are layer agnostic (decisions made per video file rather than per layer).

increases their satisfaction, and hence the potential revenue of the operator, but accelerates the consumption of network resources. Therefore, special emphasis should be placed in balancing the user experience and the network costs. The latter, in many cases, can be reduced through smart network sharing techniques.

In detail, [8] proposed a methodology which determines where to cache each video file and at what quality. These decisions are affected by the network architecture, namely the bandwidth in (and cost of) each available routing path. Besides, the caching policy is shaped by the operator which might prefer to prioritize the users' satisfaction or the suppression of its expenditures. Regarding video coding, we evaluated the impact of different encoding techniques, namely we compared versions of video files and layered encoding (SVC). We found that when user requests are heterogeneous in terms of video quality levels (i.e., users ask for both low and high quality), then SVC can be a better solution than versions.

Operators can reap higher benefits from video quality elasticity if they share their network resources. Namely, in [3] we designed caching algorithms for operators that cooperate by pooling together their physically co-located local caches,

as shown in Fig. 3(a). Instead of fetching content from distant servers, when a requested file is not available at the SCBS, the operator retrieves it from the nearby cache of another MNO. This approach allows the data to be available at the SCBS of the MNO for serving future requests besides the current. When layers are used instead of versions, there are more degrees of freedom in sharing, but the caching problem becomes more challenging.

We proposed an approximate solution to this *cooperative caching problem* by partitioning the cache capacity into portions dedicated to serve an MNO's needs and other portions for its collaborators. Using real traces of SVC-encoded videos it was shown that this solution can achieve up to 25% reduction in delay over existing layer-agnostic caching schemes (caching per video file instead of per layer) and 75% over non-cooperative caching solutions (each operator optimizes its own caching policy). As a side benefit, the proposed cooperative layer-aware caching algorithms achieve smoother playback for video streaming. The latter arise when the users view their video content while downloading it, and a key concern is to avoid video playback stalls. The cooperative layer-aware solution manages to minimize the undesirable stalls and deliver a significant portion of the requested files with very high quality, Fig. 3(b).

IV. End-to-end Video Streaming Optimization

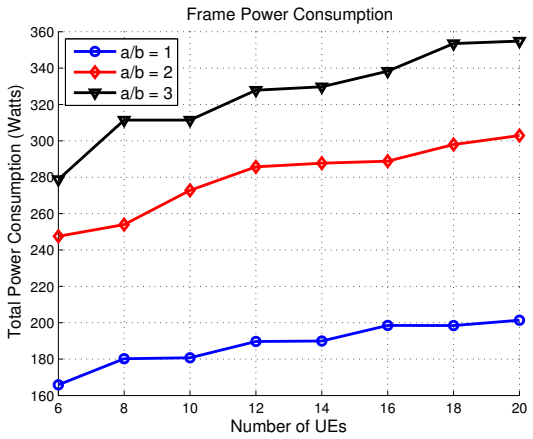
In addition to storage, operators in dense HCNs have to manage the allocation of the wireless resources like time slots, power, spectrum, and also backhaul capacity. In dense networks the challenge is that the previous tasks have to be accomplished across the numerous heterogeneous BSs in the RAN and the backhaul network infrastructure. In addition, when we consider video streaming, it is more challenging to derive the optimal resource allocation.

By performing a video-aware allocation and optimization of resources in dense networks we can have tremendous benefits for the user experience and the operator costs. However, video-aware optimization methodologies typically focus on a specific part of the network. Such an optimization may focus on the multiple BSs inside a macrocell and their multiple associated users [12], a single video streaming flow, or the storage resources in the RAN and the users [2]. However, dense HCNs will require sophisticated network decisions for the allocation of spectrum, time, backhaul capacity, the used power across different BSs, D2D resources and all this in an environment of varying channel conditions. This rather challenging and dynamic environment that emerges in the dense multi-tier architecture forces us to adopt a holistic end-to-end resource optimization approach to maximize the performance gains and minimize a potentially unbearable servicing cost.

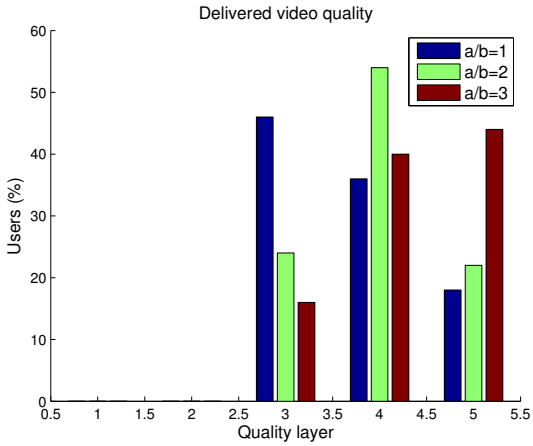
In this complex system it is essential to identify the most critical applications and focus on understanding and optimizing their performance. In [13] we focused on adaptive video streaming with DASH and considered an HCN RAN powered by LTE while we also added in the system a detailed model of the network backhaul. The first question we wanted to answer for this multi-tier system was how the backhaul network load and the total power consumption (backhaul and RAN) affect the MNO decisions for radio resource allocation, user association to the densely deployed BSs, and the final delivered video streaming quality of each user. We developed an optimization framework that employs a detailed power cost model for the entire multi-tier system (including the backhaul infrastructure). The algorithm ensures that the users are associated to the optimal BS and receive such a faction of the LTE resources that the objective is maximized. The optimization objective balances the quality of the delivered video (users' satisfaction) and the operator cost with the introduction of two balancing parameters named a and b respectively. The related results for different values of these parameters can be seen in Fig. 4(a), 4(b). It is clear that the MNO has at its arsenal a full set of operating points that tradeoff the total energy consumption and the user video quality, by carefully adjusting the parameters of the proposed framework, or alternatively by increasing the density of the multi-tier topology [13].

The second question we investigated was how to optimally design a *fast* resource allocation policy suitable for adaptive video streaming that requires only local SCBS information. We adopted a resource allocation model that corresponds to the actual operating principle of LTE-A (resource blocks and power). To ensure that the system fits in a landscape of densely deployed BS, we relax the scheduling optimization problem so that it allows video quality adaptation on a large time scale, and a fast derivation of the power and RB assignment on very short time scales. This can allow the plethora of individual BSs to reach independently a solution which, albeit suboptimal, can be employed for real dense 5G systems that have stringent time constraints.

Another question in the dense multi-tier architecture is how to manage the wireless radio resources in the complete macrocell when video streaming traffic dominates. As we explained in Section II, the multi-tier architecture of dense HCNs suffers from a specific problem, namely strong intra-cell interference. In our baseline architecture of Fig. 1 the challenge is the optimal allocation of the spectrum and time slots between the MBS and the picocells/femtocells. In [14] we investigated the implications of a dense multi-tier HCN interference on video quality. We proposed a modeling framework that can optimize the allocation of time across a macro tier and a densely deployed small cell tier, the selected video quality of the content to be delivered to each user, and the rate allocated to each user to support smooth video streaming. As the density of the small cells is increased our optimization allocates more resources to this tier by considering the video quality as the optimization metric. It is interesting to note that having a fixed allocation of the time between the different tiers leads to significantly sub-optimal performance for the video streaming users. As the video quality of the streams that the users request becomes more heterogeneous, and localized at specific BSs, optimizing the time allocated to different tiers becomes even more critical for the overall video QoE (playback freezes and video resolution).



(a) Impact of different balancing parameters on network power consumption



(b) Impact of different balancing parameters on the video quality layer that users receive

Fig. 4. Holistic resource allocation employed in the architecture of Fig. 1 (w/o storage resources).

V. Discussion and Conclusions

One of the highest priorities for the emerging ultra-dense HCNs is the efficient delivery of mobile video content. This is an equally important and challenging task. It requires the re-design of video delivery mechanisms, by taking into account the diverse types of radio and network components in emerging dense 5G systems (e.g., Base stations, backhaul links, etc.); the user demand that will be often be massive, rapidly varying, and with stringent QoE constraints; and the latest developments in video technology. The presented solutions in this paper contribute towards this ambitious goal, by providing analytical and numerical evaluation results for the design of such mechanisms.

Nevertheless, there exist many remaining issues that need to be carefully addressed. For example, the joint caching and routing (or, multicast) policies presented above, are designed based on long-term statistics, with the goal to be applied for long time periods (e.g., belong to the class of randomized algorithms). However, in certain cases the network conditions vary fast with time, e.g., links have delay that changes because of heavy interference/clutter or due to massive traffic. It is imperative to design policies that take into account such effects. For example, we need to develop methods for devising caching policies when the network links have load-dependent delay or cost parameters.

Another interesting direction is to employ more sophisticated performance objectives, i.e., to go beyond minimizing the video delivery delay. This is particularly important for video services such as streaming and video conferencing. Such objectives can include customizable criteria, such as QoE metrics, where - for example - a certain minimum video encoding level needs to be ensured for prioritized users (or, for example, sponsored content).

Finally, an equally fascinating research avenue is the development of more comprehensive tools that leverage new and sophisticated aspects of dense 5G systems and allow better understanding of video delivery performance. Besides the consideration of intra-cell interference, and backhaul costs as discussed in Section IV, additional examples include HCNs with links that operate in different bands (e.g., below 6GHz, mmWave), 5G links than span several heterogeneous BS with different hardware and PHY capabilities.

References

[1] J. Andrews, "Seven ways that hetnets are a cellular paradigm shift," *Communications, IEEE*, vol. 51, no. 3, pp. 136–144, 2013.

[2] N. Golzerai, K. Shanmugam, A. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless video content delivery through distributed caching helpers," in *Proceedings of Infocom*. IEEE, 2012.

[3] K. Poularakis, G. Iosifidis, A. Argyriou, I. Koutsopoulos, and L. Tassiulas, "Caching and operator cooperation policies for layered video content delivery," in *Proceedings of INFOCOM*. IEEE, 2016.

[4] M. Tehrani, M. Uysal, and H. Yanikomeroglu, "Device-to-device communication in 5g cellular networks: challenges, solutions, and future directions," *Communications Magazine, IEEE*, vol. 52, no. 5, pp. 86–92, May 2014.

[5] T. Stockhammer, "Dynamic adaptive streaming over http –: Standards and design principles," in *Proceedings of the Second Annual ACM Conference on Multimedia Systems*, ser. MMSys '11. New York, NY, USA: ACM, 2011, pp. 133–144. [Online]. Available: http://doi.acm.org/10.1145/1943552.1943572

[6] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the h.264/avc standard," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 17, no. 9, pp. 1103–1120, Sept 2007.

[7] F. Hartanto, J. Kangasharju, M. Reisslein, and K. W. Ross, "Caching video objects: Layers vs versions?" *Multimedia Tools Appl.*, vol. 31, no. 2, 2006.

[8] K. Poularakis, G. Iosifidis, A. Argyriou, and L. Tassiulas, "Video delivery over heterogeneous cellular networks: Optimizing cost and performance," in *Proceedings of INFOCOM*. IEEE, 2014.

[9] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5g wireless networks," *Communications Magazine, IEEE*, vol. 52, no. 8, pp. 82–89, 2014.

[10] K. Poularakis, G. Iosifidis, and L. Tassiulas, "Approximation algorithms for mobile data caching in small cell networks," *Transactions on Communications, IEEE*, vol. 62, no. 10, pp. 3665–3677, 2014.

[11] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas, "Exploiting caching and multicast for 5g wireless networks," *Transactions on Wireless Communications, IEEE*, vol. 15, no. 4, pp. 2995–3007, 2016.

[12] J. Chen, R. Mahindra, M. A. Khojastepour, S. Rangarajan, and M. Chiang, "A scheduling framework for adaptive video delivery over cellular networks," in *MobiCom*, 2013.

[13] A. Galanopoulos, G. Iosifidis, A. Argyriou, and L. Tassiulas, "Green video delivery in lte-based heterogeneous cellular networks," in *Proceedings of WoWMoM*. IEEE, 2015.

[14] A. Argyriou, D. Kosmanos, and L. Tassiulas, "Joint time-domain resource partitioning, rate allocation, and video quality adaptation in heterogeneous cellular networks," *Transactions on Multimedia, IEEE*, vol. 17, no. 5, pp. 736–745, 2015.