

Cellular-Broadcast Service Convergence through Caching for CoMP Cloud RANs

Dimitrios Christopoulos, Symeon Chatzinotas and Björn Ottersten

SnT - securityandtrust.lu, University of Luxembourg,

Email: {Symeon.Chatzinotas, Dimitrios.Christopoulos, Bjorn.Ottersten}@uni.lu

Abstract—Cellular and Broadcast services have been traditionally treated independently due to the different market requirements, thus resulting in different business models and orthogonal frequency allocations. However, with the advent of cheap memory and smart caching, this traditional paradigm can converge into a single system which can provide both services in an efficient manner. This paper focuses on multimedia delivery through an integrated network, including both a cellular (also known as unicast or broadband) and a broadcast last mile operating over shared spectrum. The subscribers of the network are equipped with a cache which can effectively create zero perceived latency for multimedia delivery, assuming that the content has been proactively and intelligently cached. The main objective of this work is to establish analytically the optimal content popularity threshold, based on a intuitive cost function. In other words, the aim is to derive which content should be broadcasted and which content should be unicasted. To facilitate this, Cooperative Multi-Point (CoMP) joint processing algorithms are employed for the uni and broad-cast PHY transmissions. To practically implement this, the integrated network controller is assumed to have access to traffic statistics in terms of content popularity. Simulation results are provided to assess the gain in terms of total spectral efficiency. A conventional system, where the two networks operate independently, is used as benchmark.

I. INTRODUCTION

A. Where it all began

In the history of wireless communications, there has been a number of services which have met commercial success and have driven not only the deployment of wireless infrastructure but also future evolutions in this domain. One of the first such services was audio and video broadcasting. The combination of low carrier frequencies with high penetration and the fact that a large number of subscribers can be simultaneously served led to ubiquitous adoption of broadcasting services. As a result, a wide range of prime spectrum is allocated to broadcasting. A proportionally sized infrastructure, in

form of radio towers and user equipment, is currently deployed worldwide.

In parallel, the past three decades a new wireless service has been constantly expanding as a response to the demand for interactive services. Although this service has evolved through a series of generations, the term cellular will be used herein as a reference name. The cellular service was originally established for bidirectional voice communications. However, during the last decade, the rapid expansion of all-IP communications is becoming the primary service delivered by cellular networks, by aggregating voice and data/Internet communications under the same umbrella. Due to the wide range of services and the advent of smartphones, cellular has met an unprecedented adoption and the design of its fifth generation network (5G) is underway. An expanding base of allocated spectrum is a substantial enabler of this upcoming new generation of cellular systems.

A key role for 5G systems is reserved by multi-antenna wireless communications. The cooperation of multiple antennas is admitted as the most prominent way of enhancing the PHY layer of wireless systems. An example of this basic concept is given in Fig. 1. An added benefit of this architecture lies in the inherent flexibility to operate the multi-antenna transmitters in either unicast or broadcast mode, which enables the considerations hereafter presented.

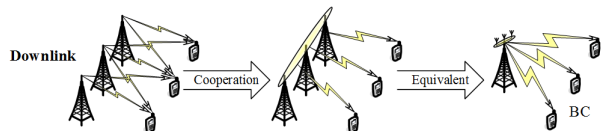


Fig. 1. Downlink base-station cooperation of a CoMP network.

A review of technical and architectural solutions that have realistic possibility to achieve the targets of future wireless access set in 5G can be found in [1]. Therein, it was argued that although further improvements in the PHY layer are expected, it is unlikely that this alone

could provide for the projected 1000-fold increase in data traffic between 2010 and 2020. In light of this conjecture, the present work aims at establishing viable and cost efficient solutions for the next generation of wireless communication systems.

B. Cloud Radio Access Networks

Included in the most disruptive technology directions for 5G [2], radio access networks organized in a cloud architecture is a recent design paradigm [3]. In the present section, we discuss a number of issues that can determine the architecture of Cloud Radio Access Networks (CRAN) [4] and highlight the main barriers and opportunities.

1) *Network Neutrality and Business Models*: Much of the Internet success actually resulted from the fact that the network is agnostic in terms of the content that it carries flexibility. On the downside, network operators have to support data hungry users without generating additional revenue. This situation has led to extremely high network throughput requirements for 5G, mainly generated by multimedia traffic combined with sub-optimal spectrum allocation. Potential solutions include additional spectrum allocation, infrastructure densification and higher spectral efficiency.

2) *Additional Spectrum*: In practice spectrum is very congested and will thus struggle to satisfy the projected throughput increase in 5G. Particularly, spectrum from 15GHz and above, is envisaged to provide short range, low complexity hot spot coverage especially for indoor environments, while lower frequencies are reserved for wider area, cellular coverage [5]. The worthiness of multicell coordination in indoor environments has been studied in [6].

3) *Infrastructure Densification*: Another approach would be to enable denser spatial frequency reuse through the deployment of micro/femto cells. Nevertheless, this would entail a large investment from the network operators which do not share the additional revenue generated by a higher-throughput network. Furthermore, an aggressive frequency reuse in a variable-cell size network creates a range of intra-system interference issues which have to be tackled before deployment.

4) *Higher Spectral Efficiency*: Improving the spectral efficiency is the holy grail of 5G. Nevertheless, Cellular systems of the previous four generations are already approaching the channel capacity bounds. Coordinated MultiPoint (CoMP), which enables full frequency reuse across the whole network seems to require further resources to cover for the projected demands.

In this context, the main question is: **How could CRANs in the context of 5G provide an even higher throughput?**

C. Evolving Traffic Patterns & Multimedia Delivery

Numerous works have focused on analyzing the traffic demand on networks delivering internet traffic [7], [8]. In order to better understand the problem we have to delve into the patterns of current Internet traffic. Originally, Internet traffic was heavily based on web browsing, which is bursty in nature and did not contain large multimedia files. However, nowadays more and more applications and services require streaming of large multimedia files and mainly video. This shift in traffic patterns is mainly motivated by the consumer habits, which have evolved from consuming linear broadcasted content to on-demand multimedia. Based on recent statistics, see for instance [9], more than half of the Internet downlink is dedicated to video streaming services, such as YouTube and NetFlix. More importantly, it seems that certain multimedia files are extremely popular and are downloaded by a large percentage of subscribers. The most prominent example of such content are the so-called viral videos. Such traffic demand patterns can be captured by a popularity distribution function. The popularity distribution of available multimedia content exhibits heavy tails, meaning that a very small number of multimedia are almost globally requested and a very large number of multimedia are almost individually consumed [10]. However, current cellular standards have been designed for unicast services, such as web browsing, and cannot exploit these evolving traffic patterns. This entails that even highly popular content would have to be unicasted to each individual subscriber by treating the cellular network as a “dump pipe”. Based on these facts, the evolved multimedia broadcast/multicast service (eMBMS) has been included in LTE-advanced [11].

D. Cellular-Broadcast Convergence through Caching

In this section, we describe a possible solution to the aforementioned problem. The key point is that globally popular multimedia can be much more efficiently distributed through broadcasting services, since it is meant to reach a high percentage of the subscriber base. Nevertheless, the subscriber expects to consume this content on demand and this implies that content should be stored locally through intelligent caching algorithms. Multimedia that are seldomly requested can be distributed through the usual unicasting-cellular services. In the following

paragraphs, we develop these ideas by providing more details on specific aspects:

1) *Cellular-Broadcast Convergence*: As mentioned before, these two modes of multimedia distribution currently form two independent networks, which follow their own standards and use separate frequency ranges. It is argued herein that there is much to be gained in terms of spectral efficiency by considering converging cellular and broadcast services. Firstly, the available spectrum can be flexibly integrated and optimized to match the evolving traffic patterns. Secondly, broadcasters and multimedia providers would be able to benefit from the cellular uplink in order to determine and follow the current traffic trends. Finally, the multimedia delivery could be optimized by selecting the cellular or broadcast mode in order to optimize the system throughput.

2) *Proactive Intelligent Caching*: This feature is a highly critical part of the solution, since it enables the subscriber to access the requested multimedia with zero-perceived delay. Caching looks more feasible nowadays, since the memory cost per GB keeps dropping. In this direction, including a large cache at the subscriber equipment is becoming a financially viable solution. At the same time, many works (eg. [12] and the references therein) have investigated intelligent caching algorithms which aim at optimizing the caching and updating actions based on a number of objectives, such as global/local popularity [13], cache/file size and temporal variations.

In this paper, we provide an analytical model for evaluating the system throughput gain of the 5G proposed architecture in comparison to conventional cellular or broadcast architectures. For the sake of fairness, this comparison considers the same amount of available bandwidth and radio towers and focuses solely on improving the spectral efficiency by minimizing the time required to deliver a fixed data volume to the subscribers.

The remainder of this paper is structured as follows: Section II presents the system model while Sec. III describes the considered multimedia traffic model and investigates the optimization of mode selection. In Sec. IV, the effect of different traffic parameters is numerically evaluated and Sec. V concludes the paper.

II. ARCHITECTURE & PHY TRANSMISSION

The present work, aims at providing a high level description of a novel concept and is therefore avoiding to accurately model an accurate wireless communication system. Instead, for the sake of simplicity, let us consider a cellular deployment of Base Stations (BSs), each being

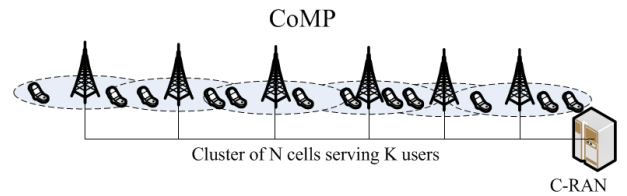


Fig. 2. Linear arrangement of a cellular system with N cooperating base stations and K users.

equipped with a single antenna. We focus on a cluster of N BSs which serve K single-antenna users in total with $K \gg N$. This cellular network can operate in two transmission modes: Unicast, where individual messages are transmitted to each user and Broadcast, where a single message is transmitted to all users. These two modes are used orthogonally, either in frequency or in time, so there is no intra-system interference between them. Furthermore, in order to optimize the spectral efficiency of the system, it is assumed that all BSs are connected to a central processor (e.g. C-RAN), which has access to both user requests and Channel State Information (CSI). This C-RAN is responsible for selecting the appropriate mode and calculating the signals to be transmitted by each BS.

To mathematically model our system, the received signal at the i -th user will read as $y_i = \mathbf{h}_i^T \mathbf{x} + n_i$, where \mathbf{h}_i^T is a $1 \times N$ vector composed of the channel coefficients (i.e. channel gains and phases) between the i -th user and the N BSs, \mathbf{x} is the $N \times 1$ vector of the transmitted symbols and n_i is the complex circular symmetric (c.c.s.) independent identically distributed (i.i.d) zero mean Additive White Gaussian Noise (AWGN), measured at the i -th user's receiver. Let $\mathbf{w}_k \in \mathbb{C}^{N \times 1}$ denote the precoding weight vector applied to the Distributed Antenna System (DAS) to beamform towards all users. In the unicast mode, the assumption of independent data transmitted to different groups renders the symbol streams $\{s_k\}_{k=1}^K$ mutually uncorrelated. The power radiated by each BS is a linear combination of all precoders [14], $P_n = \left[\sum_{k=1}^K \mathbf{w}_k \mathbf{w}_k^\dagger \right]_{nn}$, where n is the antenna index. The general linear signal model in vector form reads as $\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n} = \mathbf{H}\mathbf{W}\mathbf{s} + \mathbf{n}$, where \mathbf{y} and $\mathbf{n} \in \mathbb{C}^K$ and $\mathbf{x} \in \mathbb{C}^N$. The channel matrix $\mathbf{H} \in \mathbb{C}^{K \times N}$ is composed of coefficients dependent on the assumed channel model. By collecting all precoding vectors, a total precoding matrix $\mathbf{W} \in \mathbb{C}^{N \times N}$ is realized. In the unicast mode, since only linear precoding methods are considered, $K = N$ users will be simultaneously served. On the other hand, in the multicast mode, more

users than antennas can be served since a common message is transmitted to all users. Thus a multicast model is realised [15].

A. Operation Modes

1) *Unicast Mode (UC)*: In this mode, the BS cluster transmits an individual stream to each user and thus it can support only N users simultaneously. In this context, the max-min fair SINR optimization problem [16] with per-BS power constraint is defined as:

$$\begin{aligned} UC : \quad & \max_{t, \{\mathbf{w}_k\}_{k=1}^N} t \\ \text{s. t.} \quad & \frac{|\mathbf{w}_k^\dagger \mathbf{h}_i|^2}{\sum_{l \neq k} |\mathbf{w}_l^\dagger \mathbf{h}_i|^2 + \sigma_i^2} \geq t, \\ & \forall i, l, k, \in \{1 \dots N\}, \end{aligned} \quad (1)$$

$$\text{a.t.} \quad \left[\sum_{k=1}^N \mathbf{w}_k \mathbf{w}_k^\dagger \right]_{nn} \leq P_n, \forall n \in \{1 \dots N\} \quad (2)$$

where t is the optimization slack variable. The solution of this problem can provide the per-user spectral efficiency for the unicast mode spf_{uni} .

2) *Broadcast Mode (BC)*: In this mode, a single stream is transmitted to all K users and the caching algorithm of each user decides which content should be cached for future consumption based on the individual user requests.

Considering practical, per-BS power constraints, the max-min fair, multicast beamforming problem reads as

$$\begin{aligned} BC : \quad & \max_{t, \mathbf{w}} t \\ \text{subject to} \quad & \frac{|\mathbf{w}^\dagger \mathbf{h}_i|^2}{\sigma_i^2} \geq t, \quad i = 1 \dots K, \end{aligned} \quad (3)$$

$$\text{and to} \quad \left[\mathbf{w} \mathbf{w}^\dagger \right]_{nn} \leq P_n, \forall n \in \{1 \dots N\}, \quad (4)$$

where $\mathbf{w}_k \in \mathbb{C}^N$ and $t \in \mathbb{R}^+$. This problem is an instance of the weighted fair multigroup multicast beamforming problem of [16], for one multicast group and for equal weights. The latter constraints come from the broadcasting assumption, where a single transmission modulation and code rate configuration needs to apply for all users. More details for the solution of this problem are provided therein. The solution of this problem can provide the per-user spectral efficiency for the broadcast mode spf_{bc} .

III. MULTIMEDIA TRAFFIC MODEL

In order to study this problem, a traffic model based on the multimedia popularity is defined in this section. The popularity is measured based on the number of requests

and can be tractably described though a probability function. A widely-used abstraction for this function is the Zipf law, which is given by [10]

$$f(i) = \left(\frac{1}{i} \right)^\alpha \quad (5)$$

In more detail, if we ordered the files from most to least popular at a given point in time, then the relationship governing the frequency at which the file of rank i will appear is given by (5). Consequently, the probability of a request occurring for file i is inversely proportional to its rank, with a shaping parameter α . A detailed analysis on how to choose the exact value of α is part of future extensions of this work [17]. For the sake of simplicity, it is assumed that all multimedia files have equal size and that there are no temporal variations of the popularity measure. In addition, we assume that the cache size is not a limiting factor in this study and relaxations of these constraints are part of future work.

A. Threshold Optimization

To easily present the novel concepts of the present work, the spectral efficiencies of the previous section will be employed as deterministic parameters to assist the optimization process. This can be achieved by defining a popularity threshold below which the multimedia content is broadcasted and above which, is unicast. The decision on which mode to operate will be made based on the demand of the i -th file. In order to determine the optimal threshold above which all packets will be unicasted, let us proceed with the following assumptions. Each of the K users requires only one file of size s [bits] out of i_{max} files. Let i_{th} denote the optimal threshold and spf_{bc} and spf_{uni} denote the spectral efficiencies of the BC and the unicast configurations respectively, while W denote the total available bandwidth. Then the transmitted volume of data in BC mode will be given by $V_{\text{bc}} = (i_{\text{th}} - 1) \cdot s$, since each file has to be broadcasted only once. The received/cached data volume through the BC mode across all K users will be $s \cdot K \cdot \sum_{i=1}^{i_{\text{th}}-1} f(i)$, because the function f determines the percentage of users that have requested the broadcasted files. On the other hand, in unicast-cellular mode, the transmitted volume of data will be $V_{\text{uni}} = s \cdot K \cdot \sum_{i=i_{\text{th}}}^{i_{\text{max}}} f(i)$, since each file is individually transmitted to each user that was requested it. As a result, the received volume is by definition equal to the transmitted volume. Assuming a fixed operational bandwidth W , the goal is to find the optimal threshold

i_{th} that minimizes the total transmission time¹, thus maximizing the spectral efficiency of the total system:

$$\mathcal{T} : \min_i T_{tot} = T_{uni} + T_{bc} \quad (6)$$

where $T_{uni} = V_{uni}/(W \cdot spf_{uni})$ and $T_{bc} = V_{bc}/(W \cdot spf_{bc})$. For $\alpha > 1$ in (5), the optimization problem (6) can be straightforwardly solved by derivation. The optimal threshold between broadcasting and unicasting information in the cellular network is given by optimizing the following cost function

$$T_{tot}(i_{th}) = s/W \cdot \left(K \frac{\sum_{i=i_{th}}^{i_{max}} f(i)}{C_{max} spf_{uni}} + \frac{i_{th} - 1}{spf_{bc}} \right) \quad (7)$$

where $C_{max} = \sum_{i=1}^{i_{max}} f(i)$ is a normalization parameter for the discretized pdf f and i_{max} is the last available packet index. T_{tot} is a linear function over i . Following tedious algebraic derivations and by replacing the summations over discrete functions with integrals, i.e. $\sum_{i=1}^{i_{th}} f(i) = \int_{i=1}^{i_{th}} f(t)dt$, the optimal threshold between broadcasting and unicasting data is given by

$$i_{th} = \left(\frac{K \cdot spf_{bc}}{spf_{uni} C_{max} + K \cdot spf_{bc} \cdot f(i_{max})} \right)^{\frac{1}{\alpha}}, \quad (8)$$

and the optimal value is derived for $T_{tot}(i_{th})$.

From (8) we can see that the optimal threshold between broadcasting and unicasting data depends on the spectral efficiencies of the two systems, the total number of users and the parameters of the traffic pdf. From this expression, some conclusions can be straightforwardly reached: 1) Increasing the number of system users or the BC spectral efficiency favors the BC mode, 2) Increasing the UC spectral efficiency or the shaping parameter favours the UC mode. The latter finding can be intuitively explained by the fact that higher shaping parameters generate higher peaks and longer tails in the traffic popularity distribution.

IV. NUMERICAL RESULTS

To the end of establishing two representative values for spf_{bc} and spf_{uni} let us consider a linear arrangement of BSs, according to the well-known Wyner model [18]. Over these cells, random uniformly distributed users are generated. Signals propagate following an exponential path loss model with exponent $\eta = 3$. Rayleigh fading is assumed on top of the path loss model. In more detail

¹It should be noted that an equivalent formulation can be expressed in terms of fixing the required time and minimizing the required bandwidth.

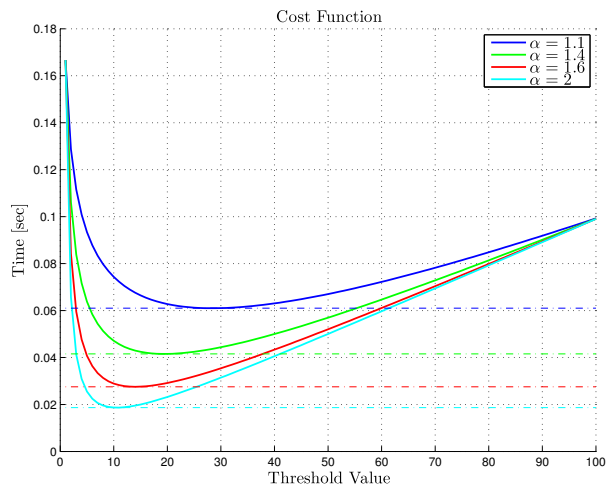


Fig. 3. Cost function versus a varying threshold value i_{th} , for various traffic parameters α . The minimum value of each function is also pointed out.

a channel coefficient between the i -th user and the j -th BS is generated as

$$h_{ij} = \frac{g_{ij}}{1 + (d_{ij})^{\eta/2}} \quad (9)$$

where $g_{ij} \sim \mathcal{CN}(0, 1)$ and d_{ij} is the distance between the i -th user and the j -th BS. By normalizing the noise power to one, the power of (2) and (4) represents the transmit signal to noise ratio (SNR). In this setup, MC simulations are performed to deduce an average spectral efficiency of the two operations, for $N = 6$ BSs, $K = 500$ system users, $i_{max} = 100$ files and $P = 1$ dBW per-BS power. In broadcast mode, the precoders are calculated by the optimization problem \mathcal{BC} . An average spectral efficiency for this system is calculated as $spf_{bc} = 1$ b/s/Hz. In the unicast mode of operation, the \mathcal{UC} optimization gives an average spectral efficiency of operation of $spf_{uc} = 3$ b/s/Hz. Consequently, the ratio of spectral efficiencies is calculated as: $spf_{uni}/spf_{bc} = 3$. Based on this setting, the cost function of problem \mathcal{T} is plotted in Fig. 3 for a range of traffic shaping parameters. It should be noted that the two extremes of this figure represent the time needed by only cellular (far left) and only broadcast (far right delivery). For $\alpha = 1.1$, which is considered a typical value [19], the data traffic along with the derived optimal point (red line) are plotted in Fig. 4.

V. CONCLUSION

In this paper, a new approach has been proposed for improving the throughput of multimedia delivery in 5G wireless networks by combining the advantages of

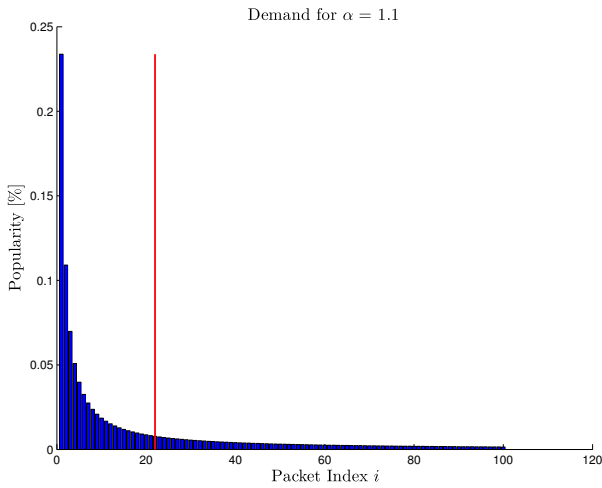


Fig. 4. Traffic demand versus packet/file index, for a traffic parameter $\alpha = 1.1$ and optimal threshold value between broadcasting and unicasting data.

unicast and broadcast transmissions. Taking into account multimedia content popularity statistics and proactive caching, highly popular content can be efficiently broadcasted with high spectral efficiency and cached by the majority of users for “virtually” on-demand consumption. On the other hand, rarely requested content is unicasted through the cellular network to the individual users for real-time consumption. In this context, an analytic framework was elaborated for determining the popularity threshold which delimits the files to be broadcasted from the files to be unicasted. In this direction, numerical results have demonstrated throughput improvements of up to 80% in comparison to conventional single-mode systems.

Acknowledgments: This work was partially supported by FNR under the project “SeMIGod”.

REFERENCES

- [1] J. Zander and P. Mahonen, “Riding the data tsunami in the cloud: myths and challenges in future wireless access,” *IEEE Commun. Mag.*, vol. 51, no. 3, pp. 145–151, Mar. 2013.
- [2] F. Boccardi, R. Heath, A. Lozano, T. Marzetta, and P. Popovski, “Five disruptive technology directions for 5g,” *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.
- [3] K. Meerja, A. Shami, and A. Refaey, “Hailing cloud empowered radio access networks,” *IEEE Trans. Wireless Commun.*, vol. 22, no. 1, pp. 122–129, Feb. 2015.
- [4] C. Mobile, “C-ran: the road towards green ran,” *White Paper*, vol. 2, 2011.
- [5] E. Dahlman, G. Mildh, S. Parkvall, J. Peisa, J. Sachs, and Y. Selén, “5G radio access,” *Ericsson Review*, 2014.
- [6] D. H. Kang, K. W. Sung, and J. Zander, “Is multicell interference coordination worthwhile in indoor wireless broadband systems?” in *GLOBECOM*. IEEE, Dec. 2012, pp. 4255–4260.
- [7] D. P. Venmani, D. Zeglache, and Y. Gourhant, “Demystifying link congestion in 4G-LTE backhaul using openflow,” in *5th NTMS*. IEEE, 2012, pp. 1–8.
- [8] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, “Understanding traffic dynamics in cellular data networks,” in *INFOCOM*. IEEE, 2011, pp. 882–890.
- [9] C. V. Forecast, “Cisco visual networking index: Global mobile data traffic forecast update 2013-2018,” *Cisco White Paper*, vol. 9, Feb. 2014.
- [10] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, “Web caching and Zipf-like distributions: Evidence and implications,” in *INFOCOM*, vol. 1. IEEE, 1999, pp. 126–134.
- [11] D. Lecompte and F. Gabin, “Evolved multimedia broadcast/multicast service (eMBMS) in LTE-advanced: overview and Rel-11 enhancements,” *Communications Magazine, IEEE*, vol. 50, no. 11, pp. 68–74, 2012.
- [12] S. Borst, V. Gupta, and A. Walid, “Distributed caching algorithms for content distribution networks,” in *INFOCOM, 2010 Proceedings IEEE*, Mar. 2010, pp. 1–9.
- [13] C. Brinton, E. Aryafar, S. Corda, S. Russo, R. Reinoso, and M. Chiang, “An intelligent satellite multicast and caching overlay for CDNs to improve performance in video applications,” Oct. 2013.
- [14] W. Yu and T. Lan, “Transmitter optimization for the multi-antenna downlink with per-antenna power constraints,” *IEEE Trans. Signal Process.*, vol. 55, no. 6, pp. 2646–2660, June 2007.
- [15] N. Sidiropoulos, T. Davidson, and Z.-Q. Luo, “Transmit beamforming for physical-layer multicasting,” *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2239–2251, 2006.
- [16] D. Christopoulos, S. Chatzinotas, and B. Ottersten, “Weighted fair multicast multigroup beamforming under per-antenna power constraints,” *IEEE Trans. Signal Process.*, vol. 62, no. 19, pp. 5132–5142, Oct. 2014.
- [17] F. Hernandez-Campos, J. Marron, G. Samorodnitsky, and F. D. Smith, “Variable heavy tails in internet traffic,” *Performance Evaluation*, vol. 58, no. 2, pp. 261–284, 2004.
- [18] A. Wyner, “Shannon-theoretic approach to a Gaussian cellular multiple-access channel,” *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 1713–1727, Nov 1994.
- [19] J. Erman, A. Gerber, M. Hajiaghayi, D. Pei, S. Sen, and O. Spatscheck, “To cache or not to cache: The 3g case,” *Internet Computing*, vol. 15, no. 2, pp. 27–34, 2011.