# Care to Share? An Empirical Analysis of Capacity Enhancement by Sharing at the Edge

Aravindh Raman
King's College London

Nishanth Sastry
King's College London

Nader Mokari
Tarbiat Modares Univ.

Mostafa Salehi
University of Tehran

Tooba Faisal
King's Collge London

Andrew Secker
BBC R&D

Jigna Chandaria
BBC R&D

## ABSTRACT

The exponential growth in online content consumption is a key concern for designing future generation network architectures. In this paper, we use content access patterns from a large trace of content accesses comprising about half the population of United Kingdom to make the case that a large portion of the backhaul load can be mitigated by content sharing amongst edge devices. We explore various models for edge devices to store and share content amongst each other, ranging from reactive opportunistic sharing to predicting future content access and speculatively placing content on strategic devices prior to request. We analyse the performance of each of these models in terms of content placement and traffic savings, which are constrained by the storage available on edge devices, the performance of the speculation engine and the wireless channel conditions. We formulate and solve at scale an optimisation problem for strategically placing content for sharing within a geographically localised cell to show such an approach can save up to 47% of the traffic generated from a small cell.

## CCS CONCEPTS

• **Networks** → **Network performance evaluation**; *Network architectures*; *Network management*;

## KEYWORDS

edge coordination, content sharing, wifi offloading

**ACM Reference format:**
Aravindh Raman, Nishanth Sastry, Nader Mokari, Mostafa Salehi, Tooba Faisal, Andrew Secker, and Jigna Chandaria. 2018. Care to Share? An Empirical Analysis of Capacity Enhancement by Sharing at the Edge. In *Proceedings of 2018 Technologies for the Wireless Edge Workshop, New Delhi, India, November 2, 2018 (EdgeTech'18: ),* 5 pages.
DOI: https://doi.org/10.1145/3266276.3266279


*EdgeTech'18: , November 2, 2018, New Delhi, India*

ACM ISBN 978-1-4503-5931-3/18/11…$15.00
DOI: https://doi.org/10.1145/3266276.3266279

## 1 INTRODUCTION

To handle the growing appetite for content, several novel *edge caching* architectures have been proposed [15, 24, 27]. The main principle behind most of these edge caching techniques is to store the content closer to the user focusing on a specific mechanism for content discovery (eg., via content-naming or HTTP/DNS redirection mechanisms) and access technology (eg., cellular or wired). Numerous suggestions have also been made to cache at several parts of the network, ranging from caching at all intermediate nodes between the requester and the serving node [5, 7] to caching selectively on the most central nodes [4] and hybrid approaches [23].

In deciding where to cache, each architecture needs to resolve a fundamental tension: Caching close to the edge keeps the content close to the user, and therefore decreases redundant transmissions of the same content in the rest of the network, freeing up the core. However, most networks are organised in a tree-like distribution hierarchy with fewer and fewer users served by distribution points closer to the edge. Thus, a cache close to the edge serves fewer users than a cache that is more central, and proactively making a copy close to the edge may prove unnecessary if no other user underneath that caching point requests the item. Hence this calls for more selective and efficient caching by intelligently choosing cache locations [4, 23] and optimising content placement [22].

We wish to turn this thinking on its head and ask whether, and to what extent, savings can be achieved by caching at the *very last hop of the network*, on a device in users' homes, and then connecting the caches together to increase sharing. This approach has several potential benefits: Developing and deploying caches on intermediate nodes in the network requires careful network planning and provisioning, as well as co-operation from ISPs, whereas caches can be deployed much more easily at the edge, even without ISP involvement. Indeed, some versions of media streaming devices such as Google Chromecast and Apple TV come with attached storage, which can serve as caches. Set-top boxes and game

consoles with huge amounts of storage are increasingly connected to Wi-Fi and the Internet.

In previous work, we explored the feasibility of such a collaborative edge architecture, termed Wi-Stitch, because it essentially involves "stitching together" a distributed content delivery network at the edge [19]. Our work showed that sharing at the edge is effective because of the so-called "chaotic" deployment of Wi-Fi networks [1], which leads to most end users being often within range of the access points of multiple neighbours. In essence, although human settlements have varying population densities, in most cases, users are "within range" of other users. Sharing among even a small number of neighbours ends up being typically effective because of the often disproportionate amount of interest on a small number of popular items.

This work extends Wi-Stitch by resolving two questions: First, since Wi-Fi AP association is limited, we ask if Wi-Stitch were deployed, will the sharing nodes have enough bandwidth to share? Our data-driven study with a TV streaming application used by the equivalent of nearly half the UK population shows that in the small cells formed by Wi-Fi sharing, simultaneous accesses to the same content and cache instance are infrequent, implying that Wi-Stitch can effectively serve content which can be cached.

This leads naturally to the second question: Wi-Stitch operates with a distributed storage capacity on nodes placed in homes. Opportunistically caching content watched by each user and sharing it out to neighbours who need it may result in multiple copies of highly popular content, and not enough copies of less popular items. We show that with an Oracle, by coordinating on which node to place each content item, we can potentially save nearly 45% of traffic in dense urban metro settings. In real deployments, Oracles will be replaced with predictive models (e.g., [8, 9]) to decide what node will access which content items, and the performance depends on the accuracy of such models. We show, with current models which can achieve accuracy of up to 90% [8], performance can approach within 10% of the savings obtained by the Oracle.

## 2 RELATED WORK

Given the exponential growth of video and other rich media traffic, a number of proposals have been made to mitigate their impact, using caching. Caching plays a central role in the design of content delivery networks [13, 16], which underpin many of today's video delivery platforms, as well as proposals to offload mobile data by storing temporarily until required by the mobile device [6, 11]. Likewise, caching is expected to play a large role in future network proposals such as 5G, which have stringent requirements on latency, bandwidth, etc., that can only be met by ubiquitous and proactive caching [3, 25].

Traditional wisdom is that caches need only to be populated reactively: caching an item on first access provides benefits to all other subsequent users requesting the same content, and doing this proactively before first access (rather than reactively) runs the risk of becoming a costly and unnecessary action if the predicted cache access does not materialise. Because our proposal relies on distributed caches, it needs to solve the content placement problem of what content to cache *where*. This has received considerable attention in literature [2, 17, 22, 26]. Whereas our approach solves one particular optimisation problem (i.e, maximising traffic savings), it would be interesting to combine some of these alternate approaches, if other optimisation goals are important.

This paper follows a line of work looking at traffic and energy savings for BBC content accesses. For instance, [10] looked at factors that affect nationwide take up of the BBC iPlayer streaming application. [9] used P2P swarms within each ISP to offload traffic from the content provider's server (but not the ISP), showcasing both traffic and energy benefits [18]. Nencioni *et al.* [14] uses set-top boxes to speculatively record content for future access, and completely offloads requests for such content from the network. Following this, Wi-Stitch [19], introduced the notion and established the feasibility of content-sharing amongst edge devices. In this work, we explore further and look at two technical aspects – whether the technology can support such a demand, and to what extent centralized and decentralised implementations along with coordinated and uncoordinated content placement approaches benefit the content sharing within a cell.

## 3 WILL WI-FI SHARING SCALE?

In this section we give a brief overview of the datasets containing a month-long access to BBC iPlayer and explore whether such a request workload can be supported over time for content sharing over Wi-Fi.

Our data reported the equivalent of over 40% of the UK's population accessing the iPlayer during July 2014 [19]. We combine this dataset of recorded accesses to TV streaming application with a wardriving data from WiGLE[1] which provides the exact latitude and longitude of WiFi access points across the UK. As a case study, we focus on sharing possibilities among customers of British Telecom (BT), one of top nationwide ISPs operating in the UK. To understand the spectrum of sharing opportunities, we look at six administrative districts[2] of diverse population densities, ranging from Hammersmith and Fulham in London, one of the 10 most densely populated areas in the country with a population density of more than 10,000 people per square kilometre, to Eden, which has the least population density in all of the UK.

[1] https://wigle.net

[2] https://en.wikipedia.org/wiki/List_of_English_districts_by_population_density

The districts are grouped into cells of ≈100m radius, which is conservatively sufficient for accessing neighbouring WiFi access points within the cell. Note that our iPlayer data is anonymised, and while location information of the accesses exist (computed by IP geolocation, down to post code or borough level resolution), this does not give the exact latitude and longitude of the access. Thus, we map the location of iPlayer requests to WiGLE access points within that location at random and use this combined dataset in the rest of the paper. We have checked that the results reported here are robust regardless of the exact mapping, by verifying the results are consistent over 50 different random mappings.

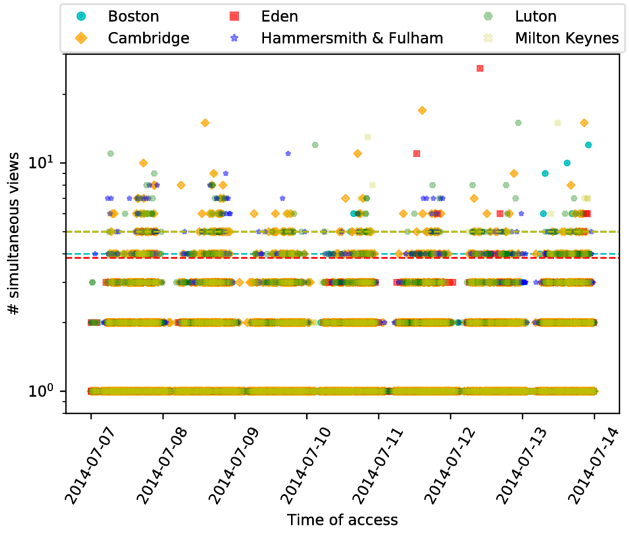


**Figure 1: Simultaneous accesses across the location over a week (at 2 minute interval), the lines indicating the $98^{th}$ percentile**

While [19] indicated a large potential for traffic savings due to shared interests and sufficient neighbours, in order to realise the traffic savings, it is equally important that the sharing infrastructure is able to support the peak demand at a time. Peak demand is captured in Fig. 1 for a random 2K people (more than the max number of users in any given cell of users who can access each other over Wi-Fi), across the six districts. Note that on any given time, *maximum* number of connections is typically no more 3–5 across the cell, and in the worst case, is only about 27 simultaneous connections surprisingly at Eden. We conjecture that this is a result of the on-demand nature of access which spreads load over time, leading to lower peak loads [14]. Thus each cell typically has to support only a handful of simultaneous connections, well within the reach of Wi-Fi.

# 4 OPTIMISING CONTENT PLACEMENT

Thus, given that even the peak demand can be handled through caching and sharing, boosting the traffic savings, in this section we look into how efficiently this content caching and sharing can be done.

The content can stored in a centralised or decentralised fashion and the content placement for this can happen through coordinated or uncoordinated approaches. In a centralised cache, there is a single cache for the entire cell (e.g., a cache on the base station in a traditional cellular setup). In the decentralised cache, storage is distributed on all the nodes within a cell. With a coordinated approach, a decision is made on where (i.e., which node) to store each content item, based on storage constraints on all the nodes, reachability (quality of wireless link) to its neighbours, and probability that its neighbours access that item. The uncoordinated approach caches data reactively: For each content accessed, the accessing node first attempts to find it on neighbouring caches, and accesses from the origin server if not found locally. This is then cached locally if storage is available on the node, and then made available to neighbours. We first formulate a optimisation problem for managing coordinated content placement in a decentralised implementation and contrast with various other content placement schemes.

## 4.1 Formulation

To have optimal content placement for minimal traffic flow at the backhaul, we consider the two main parameters, (*i*) storage within a node and (*ii*) resource availability between the nodes that share the content. Traffic at the backhaul can be given by:

$$T_{BH} = T_1 + T_2 \tag{1}$$

where $T_1$ is the traffic consumed due to the mandatory first access to the content $c$ from a repository $C$. $T_2$ is the traffic consumed when the content is available with any of neighbours but cannot be fetched from the neighbour due to lack of resource (eg., sub-channel) availability. $T_1$ and $T_2$ for a cell containing $N$ edge devices can be given by:

$$T_1 = \sum_{i,c} x_{ic}|c|, \ i \in N, c \in C \tag{2}$$

$|c|$ indicates the size of the content and $x_{ic}$ is a binary decision variable that represents whether to store the content $c$ at node $i$.

$$T_2 = \sum_{i,c} (1 - \nu_{ic})\pi_{ic}|c| \tag{3}$$

where $\pi_{ic}$ indicates the probability of content $c$ being watched by $i$ and directly depends on the accuracy of prediction technique used. $\nu_{ic}$ is a binary variable indicates whether the content is available locally (i.e., on node $i$ or any of its neighbours). This is calculated as follows:

$$v_{ic} = \begin{cases} 1, & \text{if } \sum_{j \in N} x_{jc}(1 - p_{ji}^{outage}) \geq 1 \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

where $p_{ji}^{outage}$ is the outage probability of link between $i$ and $j$. The outage probability depends on the topology being used to share the content and is calculated as a probability of whether achievable capacity $C_{ij}$ between $j$ and $i$ is sufficient to satisfy the rate requirements $r_{ij}$ for content sharing within them. i.e,

$$p_{ij}^{\text{outage}} = \Pr\left\{C_{ij} \leq r_{ij}\right\} \quad (5)$$

The achievable capacity is obtained as follows:

$$C_{ij} = \sum_{n \in \Omega} B^n \log(1 + (p_{ij}^n h_{ij}^n / \sigma_{ij}^2)) \quad (6)$$

and $B^n$ is the bandwidth of the $n^{th}$ sub-channel, $\Omega$ is the set of sub-channels allocated to link $(i, j)$, $p_{ij}^n$ is the transmit power of transmitter node $i$ to the receiver node $j$ on sub-channel $n$, and $h_{ij}^n$ is the sub-channel power gain between node $i$ and $j$ on sub-channel $n$.

Since the sub-channel power gain of each link on the allocated channels are i.i.d and have exponential distributions, the outage probability for each link (i, j) is calculated as follows

$$P_{ij}^{\text{outage}} = \prod_{n \in \Omega} \int_0^{(2^{r_{ij}/B^n})-1} f_z(z)dz \quad (7)$$

$$\text{where} \quad f_z(z) = (\sigma_{ij}^2 / p_{ij}^n \mu_{ij}) \exp(-\sigma_{ij}^2 z / p_{ij}^n \mu_{ij}) \quad (8)$$

and $\mu_{ij}$ is the average channel gain as $\mu_{ij} = s_{ij}(d_{ij}/d_0)^{-\gamma_{ij}}$ where $d_{ij}$ is the distance between transmitter $i$ and receiver $j$, $d_0$ is the reference distance, $\gamma_{ij}$ is the amplitude pathloss exponent, and $s_{ij}$ characterizes the shadowing effect.

Thus, optimisation problem for minimizing the traffic at the backhaul can be given by,

$$\min_{\boldsymbol{x}} \quad T_{BH}, \quad (9a)$$

$$\text{subject to} \quad \sum_{c} x_{ic}|c| \leq S_i, \ \forall i \quad (9b)$$

where content storage management in each node is taken care by the constraint on $S_i$ (the storage at node $i$). Thus, each content $c$ is optimally placed in a node $i$ based on decision variable $x_{ic}$ in such a way to minimize $T_{BH}$.

## 4.2 Evaluation

To understand the potential benefits of the content sharing through a coordinated placement (equation 9), we compare the traffic savings with that of an uncoordinated placement and across centralized (eg., [21]) and decentralized approaches.

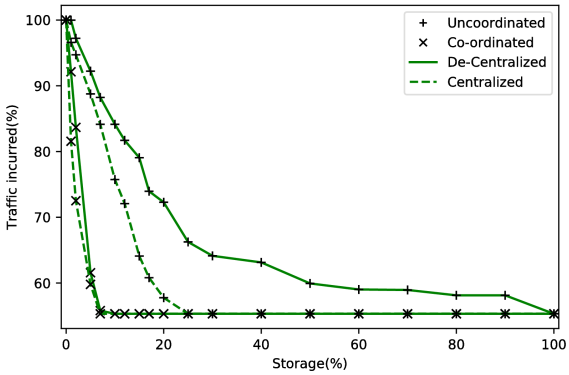


**Figure 2: Traffic savings for various combinations of operation {centralized, decentralized} x {coordinated, uncoordinated}.**

**Setup:** In order to solve equation 9 we consider the BBC iPlayer data which contains the user access to the videos, comprising mostly of catch-up TV shows mapped to the real-world location of the edge devices as explained in Section 3. This ensures content-related parameters such as size of the video (the storage of the node), bitrate at which the video is streamed (the capacity requirements between the nodes) are extracted from real-world settings. Apart from assumptions on wireless conditions for sharing described in Table 1, for evaluating $p_{ij}^{\text{outage}}$ (equation 7), we use wigle data to calculate the distance between the nodes.

Fig. 2 shows the traffic incurred based on the storage level across each of the nodes for centralized and decentralized mechanisms. It is quite clear from the figure that at lower storage levels the coordinated placement out performs the uncoordinated approach for both centralized and decentralized approaches. This also validates that our optimisation problem tends to achieve the upper bound of savings –the optimum traffic saving of ≈40% can be achieved even at a lower combined store of ≈10%.

| Parameter | Value |
|---|---|
| Total Bandwidth ($B$) | 22 MHz |
| Average path loss | 35.3 + 37.6 log($d_{ij}$) |
| Fading model | Rayleigh |
| Total number of channels | 14 |
| Shadowing | Log-normal ($\mu$=0-dB, $\sigma$=8-dB) |
| Total Transmission power | 20 dBm |
| Background noise power | 1 |
| Distance between the nodes ($d_{ij}$) | 15m - 150m |

**Table 1: Parameters used during the simulation**

Fig. 2 also shows that even with uncoordinated (reactive) content caching we can achieve a close to optimal traffic savings with an additional store of 10% - 15%. It is worth noting that the coordinated approach also depends on the prediction capability of the node. We plot the savings from a coordinated decentralized cache setting with respect to the prediction power by adjusting the accuracy of the content

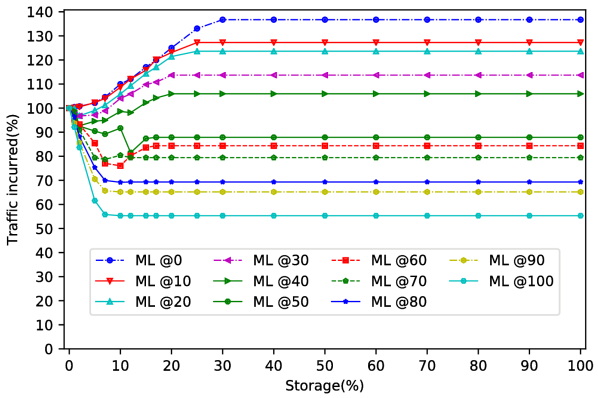


**Figure 3: Traffic incurred wrt the availability of sub-channels and the performance of predictor within a cell of 94 users.**

prediction algorithm. A predictor with accuracy lower than 50% leads to spurious cache storage decisions that actually increases traffic over the baseline of no caches. At higher levels of accuracy, performance edges closer to the Oracle. State of the art content access prediction algorithms reach accuracies of nearly 95% [8]; thus we may expect close to Oracle-like savings (within 10% of optimal).

## 5 CONCLUSION

In this work we explored the scalability of content sharing over WiFi across districts of varying population densities and formulated an optimisation problem for coordinated content sharing. We show that up to 47% traffic savings can be achieved within a small cell of ≈100 users by following this approach. We compare the combinations of coordinated and uncoordinated (reactive) content placement in a centralized and decentralized fashion and show that with an additional store of 10% to 15% a close to optimal savings can be achieved even in uncoordinated content placement. Thus using a large scale access to a popular content streaming platform and real-world distribution potential edge devices, the work evaluates the traffic savings from content sharing and validates the need for content sharing in future edge architectures. We plan to extend this further by statistically modelling the traffic savings from such cooperation using individual preference modelling to access the content [12] and broaden the content placement mechanism for the upcoming live broadcast platforms [20].

## ACKNOWLEDGMENTS

The work is partially supported by the EU-India REACH Project (ICI+/2014/342-986), by UK EPSRC via the Internet of Silicon Retinas Project (EP/P022723/1), by a gift from Vodafone and by a Professor Sir Richard Trainor scholarship.

## REFERENCES

[1] A. Akella, G. Judd, S. Seshan, and P. Steenkiste. Self-management in Chaotic Wireless Deployments. *Wireless Networks*, 13(6):737–755, 2007.
[2] D. Applegate, A. Archer, V. Gopalakrishnan, S. Lee, and K. K. Ramakrishnan. Optimal content placement for a large-scale VoD system. *IEEE/ACM Transactions on Networking*, 24(4), 2016.
[3] E. Bastug, M. Bennis, and M. Debbah. Living on the edge: The role of proactive caching in 5G wireless networks. *IEEE Communications Magazine*, 52(8), 2014.
[4] W. K. Chai, D. He, I. Psaras, and G. Pavlou. Cache "Less for More" in Information-centric Networks (extended version). *Computer Communications*, 36(7), 2013.
[5] J. Dai, Z. Hu, B. Li, J. Liu, and B. Li. Collaborative Hierarchical Caching with Dynamic Request Routing for Massive Content Distribution. In *2012 Proceedings IEEE INFOCOM*, March 2012.
[6] S. Dimatteo, P. Hui, B. Han, and V. O. Li. Cellular Traffic Offloading through WiFi Networks. In *IEEE Eighth International Conference on Mobile Ad-Hoc and Sensor Systems*, pages 192–201, 2011.
[7] V. Jacobson, M. Mosko, D. Smetters, and J. Garcia-Luna-Aceves. Content-centric networking. *Whitepaper, Palo Alto Research Center*, pages 2–4, 2007.
[8] D. Karamshuk, N. Sastry, M. Al-Bassam, A. Secker, and J. Chandaria. Take-Away TV: Recharging Work Commutes With Predictive Preloading of Catch-Up TV Content. *IEEE Journal on Selected Areas in Communications*, 34(8), Aug 2016.
[9] D. Karamshuk, N. Sastry, A. Secker, and J. Chandaria. ISP-friendly Peer-assisted On-demand Streaming of Long Duration Content in BBC iPlayer. In *IEEE Conference on Computer Communications (INFOCOM)*, April 2015.
[10] D. Karamshuk, N. Sastry, A. Secker, and J. Chandaria. On Factors Affecting the Usage and Adoption of a Nation-wide TV Streaming Service. In *2015 INFOCOM*, pages 837–845, April 2015.
[11] K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong. Mobile data offloading: How much can wifi deliver? *IEEE/ACM Transactions on Networking*, 21(2), April 2013.
[12] M. Lee, A. F. Molisch, N. Sastry, and A. Raman. Individual Preference Probability Modeling for Video Content in Wireless Caching Networks. In *GLOBECOM 2017*, pages 1–7, Dec 2017.
[13] T. Leighton. Improving performance on the internet. *Communications of the ACM*, 52(2):44–51, 2009.
[14] G. Nencioni, N. Sastry, J. Chandaria, and J. Crowcroft. Understanding and Decreasing the Network Footprint of Catch-up TV. In *Proceedings of the 22nd International Conference on World Wide Web*, 2013.
[15] G. Nencioni, N. Sastry, G. Tyson, V. Badrinarayanan, D. Karamshuk, J. Chandaria, and J. Crowcroft. SCORE: Exploiting Global Broadcasts to Create Offline Personal Channels for On-Demand Access. *IEEE/ACM Transactions on Networking*, 24(4):2429–2442, Aug 2016.
[16] E. Nygren, R. K. Sitaraman, and J. Sun. The Akamai Network: A Platform for High-Performance Internet Applications. *ACM SIGOPS Operating Systems Review*, 44(3):2–19, 2010.
[17] P. Pantazopoulos, I. Stavrakakis, A. Passarella, and M. Conti. Efficient social-aware content placement in opportunistic networks. In *Wireless On-demand Network Systems and Services, 2010*, pages 17–24. IEEE, 2010.
[18] A. Raman, D. Karamshuk, N. Sastry, A. Secker, and J. Chandaria. Consume Local: Towards Carbon Free Content Delivery. In *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*, pages 994–1003, July 2018.
[19] A. Raman, N. Sastry, A. Sathiaseelan, J. Chandaria, and A. Secker. Wi-Stitch: Content Delivery in Converged Edge Networks. In *Proceedings of the Workshop on Mobile Edge Communications*, MECOMM '17. ACM, 2017.
[20] A. Raman, G. Tyson, and N. Sastry. Facebook (A) Live?: Are Live Social Broadcasts Really Broadcasts? In *Proceedings of the World Wide Web Conference*, 2018.
[21] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire. Femtocaching: Wireless content delivery through distributed caching helpers. *IEEE Transactions on Information Theory*, 59(12):8402–8413, 2013.
[22] B. Tan and L. Massoulié. Optimal Content Placement for Peer-to-Peer Video-on-Demand systems. *IEEE/ACM Transactions on Networking*, 21(2):566–579, 2013.
[23] L. Velasco, L. M. Contreras, G. Ferraris, A. Stavdas, F. Cugini, M. Wiegand, and J. P. Fernandez-Palacios. A service-oriented hybrid access network and clouds architecture. *IEEE Communications Magazine*, 53(4):159–165, April 2015.
[24] S. Wang, X. Zhang, Y. Zhang, L. Wang, J. Yang, and W. Wang. A Survey on Mobile Edge Networks: Convergence of Computing, Caching and Communications. *IEEE Access*, 5:6757–6779, 2017.
[25] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. M. Leung. Cache in the Air: Exploiting content caching and delivery techniques for 5G systems. *IEEE Communications Magazine*, February 2014.
[26] Y. Wang, Z. Li, G. Tyson, S. Uhlig, and G. Xie. Design and evaluation of the optimal cache allocation for content-centric networking. *IEEE Transactions on Computers*, 65(1):95–107, 2016.
[27] G. Xylomenos, C. N. Ververidis, V. A. Siris, N. Fotiou, C. Tsilopoulos, X. Vasilakos, K. V. Katsaros, and G. C. Polyzos. A survey of information-centric networking research. *IEEE Communications Surveys & Tutorials*, 16(2):1024–1049, 2014.