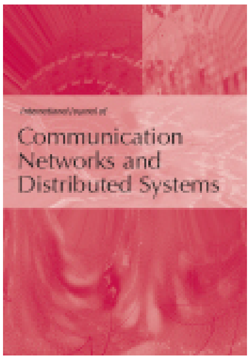
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Communication Networks and Distributed Systems

ISSN online: 1754-3924 - ISSN print: 1754-3916
https://www.inderscience.com/ijcnds

Network-aware cache provisioning and request routing in heterogeneous cellular networks

Marisangila Alves, Guilherme Piêgas Koslovski

DOI: 10.1504/IJCNDS.2025.10062283

Article History:

Received:	10 November 2023
Last revised:	12 January 2024
Accepted:	17 January 2024
Published online:	02 December 2024

Network-aware cache provisioning and request routing in heterogeneous cellular networks

Marisangila Alves and
Guilherme Piêgas Koslovski*

Graduate Program in Applied Computing,
Santa Catarina State University – UDESC,
Joinville, 89.219-710, Brazil
Email: marisangila.alves@gmail.com
Email: guilherme.koslovski@udesc.br
*Corresponding author

Abstract: In past years there has been an increase in the number of mobile devices and data traffic triggered by the popularisation of multimedia applications. In parallel, new applications and services with restricted requirements of delay and throughput have been developed. The 5G architecture is an enabler for such evolution, however, some fundamental management tasks deserve deep research, especially the cache placement. There are still challenges concerning the design of caching policies, such as limited storage space, content popularity, users requirements, and network congestion. In this context, we propose, implement, and evaluate a model of network-aware cooperative cache policy to decrease the latency experienced by heterogeneous cellular network (HCN) end-users. The model was formulated through integer linear programming (ILP) and performs both cache placement and request routing tasks. To compose a baseline for comparisons, we improved state-of-the-art policies by adding network-based attributes. Later, numerical simulations showed that the policy successfully chose paths decreasing the network latency and overcoming the counterparts.

Keywords: caching; heterogeneous cellular networks; placement; routing.

Reference to this paper should be made as follows: Alves, M. and Koslovski, G.P. (2025) ‘Network-aware cache provisioning and request routing in heterogeneous cellular networks’, *Int. J. Communication Networks and Distributed Systems*, Vol. 31, No. 1, pp.42–62.

Biographical notes: Marisangila Alves holds a Master’s degree in Applied Computing from Santa Catarina State University (UDESC) in Joinville/SC, Brazil, and an undergraduate technologist degree in Analysis and Development Systems from UDESC. She conducted research activities related to cache in Heterogeneous Cellular Networks at the Laboratory of Parallel and Distributed Processing at UDESC. Currently, she is a Professor at Santa Catarina State University (UDESC).

Guilherme Piêgas Koslovski, Professor of Computer Networks, Parallel Programming, and Distributed Systems at Santa Catarina State University (UDESC) in Joinville/SC – Brazil. He received his Doctorate from École Normale Supérieure at Lyon, France, the Master’s degree from Federal University

of Santa Maria (UFSM), Brazil, and his Bachelor's degree from the UFSM, Brazil, in Computer Science. Currently, his researches are related to high performance computing, virtual infrastructures, scheduling, software defined networks, and virtualisation of computational and communication resources. He is the Coordinator of LabP2D (Laboratory of Parallel and Distributed Processing) at UDESC.

This paper is a revised and expanded version of a paper entitled 'joint cache placement and request routing optimization in heterogeneous cellular networks' presented at *27th IEEE Symposium on Computers and Communications (ISCC)*, Rhodes, Greece, 2022.

1 Introduction

The 5G contributed to the creation of new applications and services with strict requirements of ultra low latency and high throughput (ITU, 2017). Such new requirements allow the popularisation of applications such as virtual and augmented reality, Industry 4.0, and autonomous vehicles. Specifically, the latter requirement is needed by applications based on data transfer operations and mobile video traffic (Agiwal et al., 2016).

The effective use of 5G to support these complex applications still poses network management challenges. In this sense, preventing data from travelling to the network core (e.g., cloud repositories, original databases) by delivering it directly from the network edge is a viable alternative to achieve such strict requirements. For heterogeneous cellular network (HCN) an improvement opportunity is observed, given that they implement BS with processing and storage capacity inside the mobile network backbone (Damnjanovic et al., 2011; Kamel et al., 2016), and placing data caches inside the HCN radio access network (RAN) is a promising alternative to reduce latency and replicated data transfer (Paschos et al., 2018; Kabir et al., 2020; Wu et al., 2021).

Usually, the cache-policy design addresses two main challenges: while the cache placement problem defines what, where, and how the content will be placed (Wu et al., 2021), the content delivery problem deals with requests routing and data source selection (Dehghan et al., 2017; Harutyunyan et al., 2018). The specialised literature largely focus on cache placement (related works are reviewed in Section 2.2), and regarding the requests routing, the development of cooperative cache policies is highlighted (Shanmugam et al., 2013; Bastug et al., 2014; Jiang et al., 2017). In this strategy, the total storage capacity is shared among BS, and any BS can serve as cache source for any user, independently of the original BS coverage. It is a fact that requests routing is fundamental to improve the overall performance of a networked cache system. However, the joint analysis of both problems leads to a combinatorial *NP-Hard* problem. In this context, to prune the search space, some proposals investigate only the neighbourhood of a given BS (Jiang et al., 2017), or perform a multi-hop provisioning restricting some parameters (Li et al., 2017; Song et al., 2021; Sheng et al., 2016; Xie et al., 2022). Moreover, the majority considers the network as a static entity, disregarding the real network traffic indicators and existence of other applications. *Our first claim is that cache placement and requests routing must be performed jointly, and must consider the network indicators (e.g., latency and throughput) from HCN scenarios. Consequently, all BS composing a RAN are potential candidates for* *hosting caches to deliver the strict applications requirements*. For addressing this situation, we include the round trip time (RTT) dimension while formulating the cache policy, and use the RTT evolution (increase or decrease) to obtain insights on users mobility.

Following this rationale, the cache policy must be aware that the network is shared among multiple applications, and not all applications are managed by the same operator. Specifically, this challenge requires a dynamic accounting of network links to fulfil the users' expectations. Commonly, the cache models express the total bandwidth capacity of HCN links as predefined parameters, which consequently disregard the network load evolution (i.e., traffic increase/decrease, peak loads, and eventually congestion). *Our second claim is that a network-aware cache policy must consider the dynamic aspects of an HCN.* In this sense, we largely use the consolidated literature on TCP congestion control knowledge to estimate the links and overall network load (Chiu and Jain, 1989; Brakmo and Peterson, 1995).

Given these facts, this work proposes a network-aware and cooperative policy to place caches and route users requests atop HCN BS.[1] The cache policy aims at decreasing the overall latency while guaranteeing the quality-of-service (QoS) requirements for each application. For achieving the objectives, we focus on a logically centralised management scenario, once the cache policy relies on on-going RTT to estimate network load and can dynamically perform the cache content placement atop any BS. The policy is formally modelled as an ILP, jointly addressing the capacity constraints and QoS requirements to improve the overall latency perceived by end users.

For performing a fair comparison with the specialised literature, the existing polices were classified as non-cooperative or multi-hop, and were extended to support the network dynamic aspect, as well as to receive insights from users mobility based on RTT values. This approach composed a fair baseline for comparisons, considering only network-aware models. The simulation campaign demonstrates that the aforementioned claims were successfully achieved by our cache policy in different scenarios, and showed that the policy successfully chose paths decreasing the network latency and overcoming the counterparts.

The remainder of this paper is structured as follows. Section 2 brings the background, motivation, and related work. The problem definition is presented in Section 3, while the cache policy details are given in Section 4. The numerical simulation and results are discussed in Section 5, while Section 6 presents the final remarks and future work directions.

2 Background and motivation

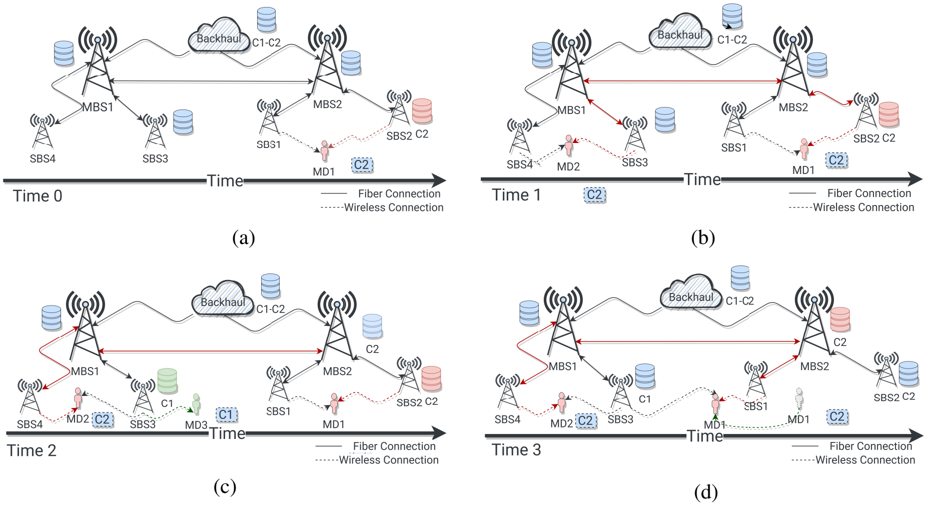
This section details the background on HCN and caches, as well as presents the related work, highlighting the research challenges and opportunities.

2.1 5G, HCN, and cache provisioning

The physical and logical proximity between resources (services, storage and computing) and end-users in HCN is essential to deliver the 5G QoS requirements (Andrews, 2013; Shanmugam et al., 2013). Specifically, the placement of caches on RAN, multi-access edge computing (MEC), and low-power nodes are natural choices to decrease end-to-end latency and increase application's throughput, as well as to reduce the replicated content load in backhaul (Pham et al., 2020; Wang et al., 2017). To exemplify this scenario, Figure 1 illustrates an HCN composed of six BS: two macro base station (MBS) and four small base station (SBS), with direct fibre connection between the MBS. Both MBS are connected to the backhaul and all BS have storage resources attached. A set of MD are utilising the caching infrastructure and services.

Initially, Figure 1(a) demonstrates the arrival of a mobile device ($MD1$), covered by $SBS1$ and $SBS2$, which requests the content $C2$, cached at $SBS2$. Later, Figure 1(b) shows a new request for the same content, originated by $MD2$. Although only covered by $SBS4$ and $SBS3$, $MD2$ can still have access to $C2$ from HCN storage resources, without requesting data from the backhaul network or creating new cache replicas. This approach relies on routing data from $SBS2$ to $SBS3$ through MBS. Following, $MD3$ requests the content $C1$ (Figure 1(c)), which is dynamically placed in $SBS3$ storage resource leading to a tricky scenario: if performed with partial data or HCN knowledge, this placement and routing decision can negatively impact the QoS perceived by $MD2$. Specifically, the network link between $MBS1$ and $SBS3$ must support the initial cache content transfer to compose the cache service at $SBS3$, as well as all cache updates, constantly competing with $MD2$ traffic. The larger the cache size and the content consumption rate, the greater the impact on other existing traffic, exemplified here by $MD2$. Consequently, a network route reconfiguration must be triggered to remove $MD2$ from such competitive scenario. Finally, Figure 1(d) depicts another challenging situation: while $MD3$ finishes the connection normally, $MD1$ is moving to another location, and the coverage signal from $SBS2$ starts to degrade, consequently decreasing the application's performance. Following, a network reoptimisation leads to a new path between $MD1$ and $C2$ (which was migrated to $MBS2$) to ensure the QoS requirements and improve the overall performance. In summary, the examples highlighted that cache placement and requests routing must be jointly performed to deliver QoS for new and running applications.

Figure 1 An example composed of mobile devices, caches, base stations and network topology: (a) time event 0; (b) time event 1; (c) time event 2 and (d) time event 3 (see online version for colours)



(a)

(b)

(c)

(d)

When scaling up the scenario from Figure 1, the applications' diversity and the network load intensify the complexity of cache management on HCN. First and foremost, the mobility of end-users on the RAN infrastructures poses a challenge regarding the dynamic routing of data between mobile devices, cache replicas eventually placed on BS, and external repositories accessed through the backhaul network. Secondly, the heterogeneity of applications requires distinct cache configurations to host multiple concurring users (e.g., a data-sharing application requires more storage cache, while a web page server may require more memory).

2.2 Related works

Content cache techniques have been a promising approach to reduce overall HCN latency and traffic on backhauls links. This section reviews the cooperative and non-cooperative network-aware cache policies, which are later used to compose the numerical simulation baseline.

2.2.1 Non-cooperative cache policies

Initially, the work (Shanmugam et al., 2013) proposed caching data on SBS (termed femtocells). The problem was formulated as an ILP with constraints related to limited storage capacity, network topology, content popularity, and latency. In turn, the work (Dehghan et al., 2017) formulated the cache problem jointly considering the routing of users' requests to optimise end-to-end latency. In summary, the cache data could be retrieved from the user's BS or directly from original repositories (e.g., cloud), and the decision process was guided by network congestion indicators and servers' load. Finally, ILP was also used in Harutyunyan et al. (2018) to jointly address cache content delivery and users' association problems to balance the network load. In addition, the authors developed an heuristic based on users' mobility aware of geographical movement.

Summarising the aforementioned works, although the users can be associated with multiple caches, once the content is not locally-available on the BS, it can be retrieved only from the remote repository. That is, there is no cooperation between caches to serve the requested content. Evidently, these policies are at a disadvantage compared to policies that use collaboration artifacts between multiple network entities to amplify the storage capacity (Li et al., 2017).

2.2.2 Cooperative cache policies

The management policy proposed by Pu et al. (2018) focus on decreasing costs on cloud radio access networks. The formulation considers constraints on storage capacity, latency, and costs for links, VM reconfiguration, and caches migration. VM and baseband units are hosted in central offices, which collaborate to implement the cache system. The cooperative model is one-hop limited. Following this line, a cooperative cache model for HCN was proposed in Jiang et al. (2017) to reduce end-to-end latency. Apart from traditional storage and communication constraints, the formulation was based on network topology details, and probability of requesting a given cache content.

The hierarchical cooperative cache provisioning was investigated in Li et al. (2017) to maximise the cache hit metric. The cache policy jointly analysed the requests routing and cache provisioning, being aware of capacity and network constraints. In such hierarchical approach, a user can connect to multiple hierarchical levels to retrieve the cache content from neighbours from distinct levels. In turn, a cooperative architecture between SBS and mobile devices was proposed by authors in Sheng et al. (2016). The multi-layer architecture decreased the overall latency by applying device-to-device content delivery. Initially, the cache is searched on local neighbourhood, and if necessary the request is recursively sent towards the topology root, characterising a multi-hop cooperation. The reliable provisioning of cache content is addressed by Song et al. (2021). The cache policy applies multi-hop cooperation, and each SBS acts a cache storage device or proxy. The work (Xie et al., 2022) proposes a multi-hop model to minimise the cache latency. Linear programming was used to implement two cache models, with and without constraints on network bandwidth.

The authors in Rafique et al. (2023) introduce a novel caching framework for latency-aware cache placement and request routing, incorporating link utilisation constraints using SDN and ICN. In turn, the work in Somesula et al. (2023) formulates an ILP for service placement. A greedy and randomised rounding technique is presented to address service caching and request routing, along with a heuristic. In the context of AR (Liu et al., 2021), the focus is on proposing caching placement and request routing aimed at viewport rendering services with lower delay. Latency and backhaul bandwidth constraints are considered. Finally, the authors in Li et al. (2024) propose Binary Integer Linear Programming and a framework that considers caching placement and user association (request routing) with the goal of minimising latency during handovers caused by mobility and re-association. In short, the cooperative cache strategies (hierarchical or multi-hop) are promising approaches compared to non-cooperative caches, as they offer better use of storage and communication resources.

2.3 Motivation and research opportunity

The related works largely focus on requests' routing problem. Moreover, the majority starts from the principle of cooperation either from a neighbouring BS, or from multi-hop/hierarchical routing. Specifically, hierarchical and one-hop policies limit the search space in terms of cache repositories and network paths, which can reduce the cache hit indicator. It is important to highlight that the specialised literature ignores the fact the HCN is shared by multiple applications, not only the cache system and mobile application. In other words, they ignore the possible network variations, which may occur in the intermediate paths (Dehghan et al., 2017; Pu et al., 2018; Jiang et al., 2017; Li et al., 2017). In turn, the multi-hop policies desconsider the devices' mobility regarding the QoS requirements (Song et al., 2021; Sheng et al., 2016; Xie et al., 2022; Rafique et al., 2023; Somesula et al., 2023; Liu et al., 2021). Furthermore, such works use bandwidth restrictions, therefore, they disregard the existence of other possible flows competing from the same link capacity, that is, the bandwidth is used as a measure of the real capacity of the link. This decision can provide ambiguous information about the state of the link and its actual capacity. In summary, the items in which the proposed network-aware cooperative cache policy differs (or improves) from the works mentioned are listed below:

- *Multi-hop cooperation and multi-path routing*: The cache content can be placed atop any BS from the HCN. Based on the multi-hop requests routing, users can have access to content from distinct sources. The routes between end devices and cache repositories are constantly optimised to improve the overall network latency perspective.
- *Actual throughput estimation*: Unlike the approaches that restrict the link capacity and, in such a way, ignore other flows that travel through the network, the proposed cache policy estimates the actual link capacity based on the TCP congestion

avoidance and control algorithms. Moreover, besides using mobility-related metrics only to define the connection between BS and users, we use them as indicators of QoS. Specifically, the policy monitors the RTT between all network pairs and paths, which can increase or decrease in function of user to BS distance. This indicator triggers a reconfiguration, once the user is moving to another BS, and consequently requires a reoptimisation of intermediate routes.

- *Network-aware model*: In general, the aforementioned works do not consider the load or congestion in the HCN FH and BH links. We opt for routing requests with a model aware of network (actual throughput) and servers load (storage capacity).

It is worthwhile to mention that in Alves and Koslovski (2022) we formulated the initial problem definition and presented calibration results regarding the network-aware cache provisioning and request routing in HCN. The simulation protocol investigated the applicability of the model, tuning the simulation parameters. In the present work we improved the ILP model and deeply compared the proposed policy with the state-of-the-art approaches. Moreover, the analysis deepens the discussion on aspects related to the network, and the results demonstrate the impact related to the latency variation. We formulate the problem and present the cache policy in the following sections.

3 Problem formulation

It is a fact that not all mobile applications are managed by the cache service providers, however those which are must detail the QoS requirements during the SLA negotiation to improve the overall performance indicators. Thus, the cache model must consider the SLA requirements specified for each application in two distinct moments:

i initially, when a new request is submitted, denoting the allocation process

ii periodically, during the applications' runtime, based on monitored data to verify the SLA concordance and eventually reconfigure the network paths and cache placement for guaranteeing the QoS constraints.

The first case (new request) is exemplified by Figure 1(a) where the mobile device $MD1$ arrives and requests the data content identified as $C2$. In turn, the second case (reconfiguration) is demonstrated by Figure 1(c): the mobile devices $MD1$, $MD2$, and $MD3$ are sharing the HCN topology and consequently disputing the networking resources. At this point, the cache policy can reoptimise the scenario based on up-to-date metrics.

The cache policy is to decide simultaneously where to cache content and how to route requests from MD to cached content. Technically, one could observe that the content should be placed in advance, later defining the appropriated routes. However, we argue that the decision-making must be jointly performed, activating the caches in appropriated places, while selecting the network-efficient routes. It is a fact that eventually the scenario must be reconfigured to accommodate new users, background traffic variation, and other network-related phenomena, however even the reconfigurations must be performed considering both the cache placement and requests routing. For this situation to be feasible and realistic, the possible locations for placing the caches must already be known by the system (e.g., the edge servers providers, MBS, and SBS), and usually, such locations are chosen in advance, based on history of use and demands (Breslau et al., 1999).

The provision of QoS-aware caches atop HCN shared by multiple applications with a high diversity on data traffic is a challenging problem. For modelling the scenario considering such dynamic load and unknown number of users, we rely on consolidated literature and algorithms from traditional end-to-end TCP congestion control, specifically on time-sensitive variants (Brakmo and Peterson, 1995; Cardwell et al., 2016). It is worthwhile to highlight that the present work considers the cache management at application layer of TCP/IP stack, however, the model employs well-defined concepts from the transport layer. Hereafter, a graph notation is used to formalise all model components and data, as summarised by Table 1.

Table 1 Formal notation to represent the problem and the cache policy

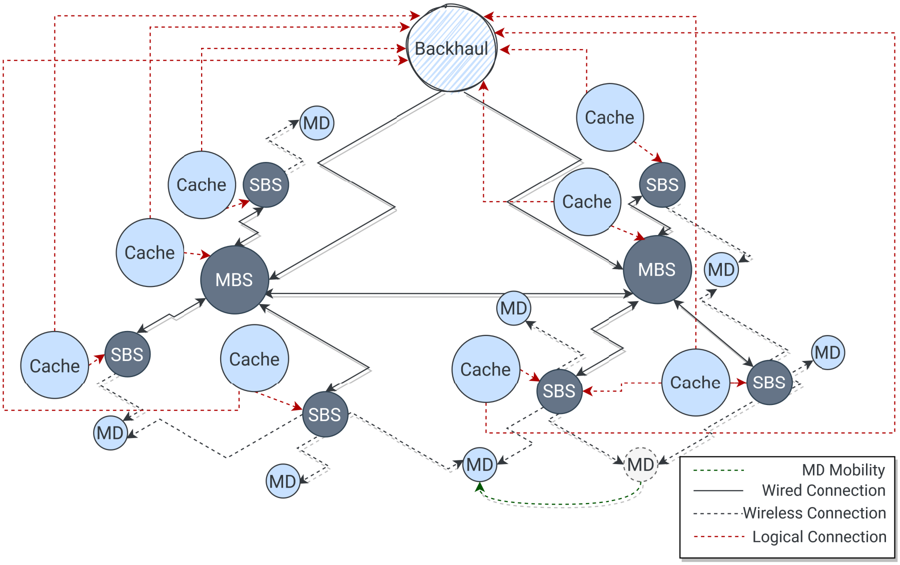
Notation	*Description*
$G(V, E)$	Graph denoting the HCN, users, cache resources, and external repositories
$V = BS \cup UE \cup C \cup S$	Vertices composing a graph: users (UE), caches (C), base stations (BS), and external repositories (S)
$E \subseteq (V \times V)$	Physical and logical connections between vertices
$c_k^s \in \mathcal{N}+$	SLA storage requirement for cache $k \in C$
$c_k^b \in \mathcal{N}+$	SLA buffer requirement for cache $k \in C$
$c_k^{thp} \in \mathcal{R}+$	SLA minimum throughput requirement for cache $k \in C$
$b_i^s \in \mathcal{N}+$	Denotes the total storage capacity for $i \in BS$
$D_i \in \mathcal{R}+$	Denotes the maximum coverage radius distance for each $i \in BS$
$\gamma_{ik} \in \{0, 1\}$	Enables the placement of a content $k \in C$ in $i \in BS$
$r_{uk} \in \{0, 1\}$	Indicates that user $u \in UE$ requested content $k \in C$
$rtt_{ij} \in \mathcal{R}+$	Latest RTT accounted for $(i, j) \in E$
$thp_{ijk} = \frac{c_k^b}{rtt_{ij}} \in \mathcal{R}+$	Latest throughput accounted in $(i, j) \in E$ for a cache $k \in C$
$x_{ijk} \in \{0, 1\}$	Denotes if the cache content $k \in C$ is flowing through a link $(i, j) \in E$
$y_{ik} \in \{0, 1\}$	Indicates the effective use of a cache service $k \in C$ hosted by $i \in BS$

3.1 Graph representation

A graph $G(V, E)$ represents the HCN, users, cache resources, and external repositories (usually the source of data that must be placed on caches - e.g., clouds, edges, and other data centers) hereafter termed backhaul. Specifically, the set of vertices V is composed of base stations (BS), representing MBS and SBS, as well as the users' mobile devices (UE), caches (C), and external repositories (S). To illustrate it, Figure 2 exemplifies the graph-based representation of Figure 1(d) scenario. We resorted to an extended graph technique (Chowdhury et al., 2012; De Souza et al., 2017; de Souza et al., 2019) to combine physical (e.g., BS, users) and logical (e.g., caches, network path, and content distribution) information into a single graph. In this sense, the solid black lines denote wired connections, while dashed grey lines represent wireless connections (the SBS coverage). In turn, dashed green lines are used to represent the users' mobility, while dashed red lines are the logical connections included by the graph extension technique, demonstrating were caches could be placed. Potential cache repositories are logically connected to backhaul (for retrieving the original data) and to BS (indicating that the BS is cache-enabled). Finally, the users mobility and BS coverage are represented by dashed green lines.

A cache service c requires $c_k^s \in \mathcal{N}+$ storage resources, a minimum buffer size defined by $c_k^b \in \mathcal{N}+$, and $c_k^{thp} \in \mathcal{R}+$ minimum end-to-end throughput to be efficiently provisioned. In turn, a base station $i \in BS$ has a total storage resource capacity donated by b_i^s (0 indicates that caching is not enabled). It is important to note that the graph composition and its attributes (residual capacities and parameters) represent a snapshot of the infrastructure. Each new snapshot will eventually contain differences from the previous one that must be considered by the cache policy. Finally, the cache policy combines the new requests with those previously allocated to perform a complete reconfiguration (addressing the cache placement and requests routing), whenever the model's parameters allow, as described as follow.

Figure 2 A graph representing users (MD), BS, external repositories, caches, and network topology. The solid black lines denote wired connections, while dashed grey lines represent the wireless connections (the SBS coverage). Dashed green lines are used to denote the users' mobility, while dashed red lines represent logical connections added to the original graph to indicate the potential cache placement using a single graph representation (see online version for colours)



3.2 *Users, caches and base stations connectivity*

The HCN topology and BS coverage radius are modeled based on geographical distances. Each BS i and user's mobile device $u \in UE$ has a pair of (x, y) coordinates associated to it, and a function $dis(.) \in \mathcal{R}+$ is applied to account for the cartesian distance between two pairs of coordinates. There is a directional connection $(i, u) \in E$ if $dis(x_i, y_i, x_u, y_u) \leq D_i$, where $D_i \in \mathcal{R}+$ denotes the maximum coverage radius distance for each BS i in the HCN scenario. In summary, an edge $(i, u) \in E$ indicates that a user's mobile device $u \in UE$ is covered by a BS i.

The connectivity between cache resources (e.g., servers or any other specialised resource (Ordonez-Lucena et al., 2017)) and BS follows the same rationale. Initially, any existing algorithm for populating caches based on content access popularity can be applied for distributing content on BS (Ayenew et al., 2021), composing the parameter γ_{ik}, which indicates that a cache k can be potentially placed on $i \in BS$. Consequently, an edge $(k, i) \in E$ denotes such initial scenario. In turn, the backhaul for retrieving the original data for any cache k is represented by a single entry point S. These two sets of logical correspondences are represented in the extended graph from Figure 2. Finally, a parameter $r_{uk} \in \{0, 1\}$ is used to indicate that a mobile device u is requiring a cache $k \in C$. It is worth mentioning that no assumption is made regarding the application-level cache configuration (e.g., device's buffer size) or the base stations handover.

3.3 Perspectives of mobile network operators and cache providers

For representing and dealing with the dynamism of HCN traffic, the model embraces the TCP congestion control knowledge and it is constructed based on RTT values. Each $(i, j) \in E$ has a RTT associated to it representing the latest sample (denoted by rtt_{ij}). Given the latest RTT, one can account the current estimated throughput (Brakmo and Peterson, 1995) for a cache $k \in C$ atop a link $(i, j) \in E$, as given by $thp_{ijk} = \frac{c_k^b}{rtt_{ij}} \in \mathcal{R}+$.

The communication mobility is a requirement for 5G and future internet scenarios. However it poses challenges in guaranteeing network-related QoS. By increasing (or even varying) the distance between devices and base stations, the quality of communication signal is directly affected and, in some cases, the mobility can result on connectivity handovers between SBS. Such facts can lead to packet losses, a factor that directly impacts the application-level RTT. It is possible to deduce that the RTT is related to the distance between the user and the SBS and furthermore, when the RTT exhibits an increasing (decreasing) trend along the time, we can infer the impact on distance (and vice-versa) (Tian et al., 2005). In turn, it is important to mention that the RTT obtained in the intermediate paths, that is, through wired connections, also vary according to the links' loads (Chiu and Jain, 1989).

To contribute to existing studies, the present work innovates by defining a network-aware cache policy based on estimation of the actual link capacity, instead of only considering the maximum link bandwidth (as it is usually an HCN classified information) (Harutyunyan et al., 2018; Pu et al., 2018; Jiang et al., 2017; Li et al., 2017; Song et al., 2021). The estimation of actual link (or path) capacity follows the end-to-end design principle of TCP congestion control algorithms enabling a feasible use in users competitive scenarios, composed of multiple services and applications. Consequently, the proposed cache policy is agnostic to concurrent traffic (e.g., applications, data transfer) on network links and paths which are not administrated by the MNO.

4 Cache policy

The cache policy's objective is to simultaneously maximise the network and cache resources usage while guaranteeing the QoS indicators for those cache providers who requested SLA requirements. The main objective is achieved by jointly performing the cache placement and the request routing. We recur to ILP to formally analyse and represent the model.

4.1 Variables and objective

The ILP relays on traditional multi-commodity flow problem for representing the network configuration, jointly considering the capacity of caches and QoS requirements defined by cache providers. In this sense, two binary variables are used: x_{ijk} denotes if the cache content k is flowing through a link $(i,j) \in E$, while y_{ik} indicates the effective use of a cache service k hosted by $i \in BS$ (by setting the value 1, and 0 otherwise). In other words, $y_{ik} = x_{kik}; \forall k \in C, i \in BS$. It is worth mentioning that equations (2)–(4) guarantee that there is a single active path between a cache and a user. To achieve the MNO' and cache providers' perspectives a minimisation-based objective function is defined by equation (1).

$$min \sum_{u \in UE} \sum_{i \in BS} \sum_{k \in C} \frac{(b_i^s - c_k^s) \cdot r_{uk} \cdot y_{ik}}{b_i^s + \delta} + \\ \sum_{u \in UE} \sum_{(i,j) \in E} \sum_{k \in C} \frac{c_k^{thp}}{thp_{ijk}} \cdot x_{ijk} \cdot r_{uk} \tag{1}$$

Both terms of the objective function aim at load balancing the demands atop the available residual resources. Although applications indicate a minimum throughput requirement (c_k^{thp}), the policy can select higher values (thp_{ijk}) based on current HCN load. In this sense, equation (1) aims at decreasing the requested-to-allocated throughput ratio to avoid congestion in the network and potentially allocating more requests. Finally, $\delta \rightarrow 0$ is a small positive constant to avoid dividing by zero when a cache is not currently offered by a given BS.

It is worth noting that the decision variables are closely related, that is, the cache placement is defined as a function of the request routing decision, and vice-versa. In addition, the first term from equation (1) aims to avoid the use of backhaul links, while the second term balances the network load giving priority to high throughput links, hence decreasing the latency.

4.2 Constraints

A set of flow- and QoS-related constraints must be satisfied while accounting the ILP objective function. Initially, the flow-related constraints are given by equations (2)–(4). While equation (2) ensures that all flows will be routed inside the graph components, equations (3) and (4) indicate that the cache data flows to the users' mobile devices (Karp, 1975). Equations (5) and (6) are used to guarantee the QoS requested by cache providers. For performing the cache placement, the model must assure that the hosting HCN components have enough storage capacity (equation 5), while the requests routing must guarantee the requested throughput (equation 6). Specifically, equation (5) accounts the BS storage capacity considering that a cache k can be concurrently accessed by multiple requests. In turn, equation 7 ensures that a requested is attended just by one cache source. This approach combined with the latency-oriented formulation (Section 3.3) aims at decreasing the backhaul pressure while placing the cache content atop fronthaul BS.

$$\sum_{i \in BS} x_{jik} - \sum_{i \in BS} x_{ijk} = 0; \forall j \in BS, \forall k \in C \tag{2}$$

$$\sum_{i \in BS} x_{kik} - \sum_{i \in BS} x_{ikk} = 1; \forall k \in C \tag{3}$$

$$\sum_{i \in BS} x_{uik} - \sum_{i \in BS} x_{iuk} = -1; \forall k \in C; \forall u \in UE \tag{4}$$

$$\sum_{u \in UE} \sum_{k \in C} c_k^s \cdot y_{ik} \cdot r_{uk} \leq b_i^s; \forall i \in BS \tag{5}$$

$$c_k^{thp} \leq thp_{ijk} \cdot x_{ijk}; \forall (i,j) \in E, \forall k \in C \tag{6}$$

$$\sum_{i \in BS} y_{ik} \cdot r_{uk} = 1; \forall k \in C, \forall u \in UE \tag{7}$$

5 Simulation and results

The discussion on related work highlighted multiple proposals to improve the QoS-aware provisioning of caches atop HCN and similar scenarios. We summarised the proposals (from Section 2.2) to compose the baseline for analysing the performance of our cache policy. In short, two major categories were prepared: the first one representing non-cooperative cache policies, and the second one based on multi-hop cooperation. However, for performing a fair comparison, we first had to improve the network perspective of both categories. In essential, the proposals must consider the dynamic aspect of HCN network. Instead of using a predefined maximum throughput capacity for each network link, we extended the proposals to gather the network load from dynamic information (i.e., RTT and estimated on-going throughput per flow).

In this section, we first describe the baseline formulations (Section 5.1), followed by details on experimental prototype and infrastructure (Section 5.2). The simulation parameters, metrics, and results as discussed in Sections 5.3–5.5, respectively.

5.1 Baseline models

According to the cache policies proposed by the specialised literature (Section 2.2), we highlighted two major strategies of baseline to analyse the performance of our network-aware cache policy, as follows:

- *Non-cooperative*: This strategy considers only one hop of forwarding atop the HCN. Therefore, the cache policy searches for the content only in BS connected to mobile device and not consider others BS in RAN.
- *Multi-hop*: This strategy searches the content in others BS in RAN. In other words, a user can retrieve the data from any BS from the HCN, instead of getting it only from the access BS.

To analyse the performance of the baselines, the fundamental concepts have been generalised, modeled and implemented as described in the following.

5.1.1 Non-cooperative model

The non-cooperative model prunes the search space by limiting the path length (number of hops) for routing a cache request (Shanmugam et al., 2013; Dehghan et al., 2017; Harutyunyan et al., 2018; Pu et al., 2018; Jiang et al., 2017; Li et al., 2017). Therefore, we add a new constraint to bound the cache's placement (equation (8)). This constraint ensures that the path between the requesting user and the cache repository is limited (two hops are needed as we modelled as an extended graph – Section 3). For enabling the use of external repositories there is only one logic hop between the BS and the origin and, another hop between the BS directly connected to the user. In addition, the network-oriented objective function (equation (1)) is adapted as given by equation (9), however the first term remains equal to perform a fair comparison. We highlight that the non-cooperative model follows the same principles of the network-aware policy, and considers the requests previously allocated when processing new requests. Thus, the baseline model also reoptimises the scenario analysing all requests. In other words, the model has a global view of the network.

$$\sum_{j \in BS} \sum_{i \in BS} \sum_{k \in C} x_{ijk} \cdot r_{uk} \leq 2; \forall u \in UE \tag{8}$$

$$min \sum_{u \in UE} \sum_{i \in BS} \sum_{k \in C} \frac{(b_i^s - c_k^s) \cdot y_{ik} \cdot r_{uk}}{b_i^s + \delta} + \sum_{u \in UE} \sum_{(i,j) \in E} \sum_{k \in C} x_{ijk} \cdot r_{uk} \tag{9}$$

5.1.2 Multi-hop model

Cache policies based on multi-hop cooperation were previously proposed (Sheng et al., 2016; Song et al., 2021; Xie et al., 2022). For comparison purposes, we modelled the multi-hop policy following the same assumptions of our network-aware model, except it does not consider the network in its optimisation function. Formally, the objective function is given by equation 9, not containing the ratio that determines the orientation to the network and mobility. The multi-hop model also performs a global optimisation, reallocating requests, if need, to accommodate new demands. Finally, the content placement nearest the user is prioritised, however, the multi-hop model is unaware of network throughput optimisation when performing the requests routing.

5.2 Prototype and infrastructure

As a proof-of-concept, we implemented all cache policies ILP models with the Gurobi Optimizer version 9.1,[2] as well as a discrete event simulator (Python 3.9). At each event, a set of cache service requests arrives to be provisioned, as well as routing optimisations are performed (based on predefined parameters). Finally, simulations were performed on a server equipped with Intel Xeon E312XX and 64 GB RAM.

5.3 Simulation parameters

The selected parameters were, whenever possible, guided by the specialised literature and are summarised by Table 2. For each comparative scenario were executed 100 discrete events (Sheng et al., 2016) with an arrival rate up to 5 new requests per event following a Poisson distribution (Dehghan et al., 2017). The HCN is composed of 2 MBS, and each MBS has 15 SBS associated with, following similar scenarios from specialised literature (Khreishah et al., 2016; Jiang et al., 2017). The SBS coverage radius is 70 meters (Shanmugam et al., 2013; Sheng et al., 2016; Jiang et al., 2017), and the storage capacity of each BS is up to 40% of the total caches sizes. A total of 200 users can connect to SBS limited by two SBS simultaneously connected per user's device (Kamel et al., 2016). The users' devices can move around randomly, as previously performed by the specialised literature (Shanmugam et al., 2013; Harutyunyan et al., 2018; Sheng et al., 2016), with steps up to 10 meters away from the current cartesian point. While a random mobility does not represent a real urban mobility scenario, it provides a fair distribution for comparison with previous work.

Each physical link has an initial RTT defined as 1 millisecond (ITU, 2017). Specifically for the one-hop model, in the edge between origin (i.e., cloud) and users devices, the initial RTT is 10 times larger the default RTT (Lyu et al., 2021). Each cache service request remains active up to 10 discrete events, and the deallocation of a request represents the finalisation of a service usage triggered by the mobile device. Regarding the quality-of-service configuration, the cache service owner specifies the buffer size and the minimum throughput, for each cache, defined as 48Mb and 100Mbps, respectively (ITU, 2017). The cache content library is composed of 100 distinct contents (Sheng et al., 2016) with sizes in 2GB, 4GB e 8GB. In turn, the γ parameter dictates the fraction of BS that can host a given cache content, configured as specified by the policy model (discussed in Section 5.1).

Table 2 Parameters for the simulation protocol. The selected parameters were, whenever possible, guided by the specialised literature

Parameter	*Value*
Number of MBS	2
Number of SBS per MBS	15
Total cache repositories (C)	100 contents
Number of mobile users (UE)	200
SBS coverage radius	70 m
Minimum application throughput (c_k^{thp})	100Mbps
BS storage capacity (b_i^s)	40$\%$ from total
Cache content size (c_k^s)	2GB/4GB/8GB
Application's buffer size (c_k^b)	48Mb
Users' mobility	10 m
Content popularity: Zipf distribution (α)	0.8
Requests arrival rate (Poisson distribution)	up to 5 requests per event
Requests duration	up to 10 events
Initial RTT	1 ms

Finally, the content popularity follows a Zipf distribution with $\alpha = 0.8$. The Zipf distribution is used to represent the users' behaviours when submitting cache requests, given as $P_k = 1/(k^{\alpha})$. Smaller values of α means that the users' interests are diversified among the contents present in the library, while larger values of α means that the users' preferences are concentrated in the same contents (Breslau et al., 1999).

5.4 Metrics

We selected three representative metrics to investigate the performance of our cache policy facing the predefined baselines: latency, cache hit, and storage usage. Initially, we define latency as the time measured in the interval between the user's request and its subsequent receipt by the cache server, that is, between origin and destination. The latency of a request is the end-to-end time interval, which in the simulations was based on the RTT counted in the logical and physical links belonging to the path taken by the request. Specifically, the RTT is the time measured between the sending of a packet and its confirmation, in this case, the RTT are counted for each link.

The cache hit represents the number of cache requests served inside the HCN. In this way, the cache hit ratio is obtained by dividing the cache hit to total requests submitted per discrete event. Finally, the total storage usage considers the total capacity of the cache system present in the RAN, that is, the sum of the capacities of each BS results in the total capacity.

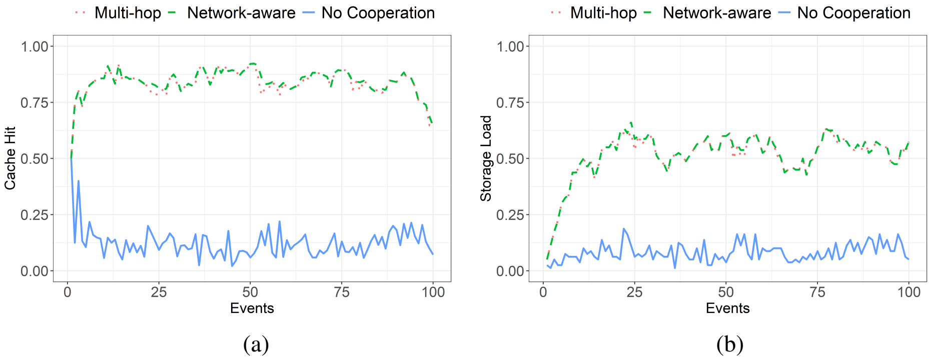
5.5 Results and discussions

The first discussion is carried out by combining the cache hit and the total HCN storage. Latter, the network aspects are discussed to corroborate our network-related claims.

5.5.1 Cache hit and HCN storage

The cache hit results for all cache models (baseline and our network-aware model) are summarised by Figure 3(a), while the total HCN storage usage is given by Figure 3(b). Initially, Figure 3(a) indicates a similar profile of cache hit for both the multi-hop and network-aware models. Specifically, both models perform the cache placement and requests routing across the entire RAN, achieving a higher cache hit ratio compared to the non-cooperative model. This behaviour is expected and reaffirms what was emphasised by (Li et al., 2017). The average cache hit for the network-aware model was 0.84 ± 0.05, while the model without cooperation achieved 0.12 ± 0.06. Finally, the multi-hop average was 0.83 ± 0.06. The average difference between the non-cooperative model and the multi-hop and network-aware models was $\approx 86\%$. In other words, both models obtained a cache hit ≈ 7 times higher compared to the non-cooperative model.

Figure 3 Cache hit and storage results: (a) cache hit comparison between all policies and (b) total HCN storage usage (see online version for colours)



It is important to highlight that although the multi-hop model considers cooperation and search for content in all BS and obtained an average similar to our network-oriented model, the approach of just placing the content without actually analysing the network does not guarantee that latency is reduced or that specific links have not been overloaded (this aspect is discussed in Section 5.5.2). Unlike the network-oriented model, the multi-hop model does not consider the network dynamics, so it does not have the ability to choose links with higher throughput in order to minimise latency.

The analysis on the total HCN storage capacity confirms the observations (Figure 3(b)). The average for the non-cooperative model was $0.08 \pm 0, 03$, while the multi-hop and network-aware obtained the same average 0.51 with standard deviation of 0.10 and 0.07, respectively. In other words, the non-cooperative model reached only 15% when compared to its counterparts. It should be noted that the resulting standard deviation is caused by the initial simulation steps, since it is considered that in the first event there was no content originally placed in cache. In summary, Figure 3(b) demonstrates that the total existing storage capacity is underutilised in the non-cooperative model.

The lower proportion of local cache hit obtained by the non-cooperative approach indicates that more content must be transferred through the BH links, which may be less advantageous from the mobile network operator (MNO) perspective, in addition to potentially increasing the latency. On the other hand, the multi-hop and network-oriented models introduce a higher load on FH links distributing the traffic along the RAN components. In general, the multi-hop and the network-oriented models prioritise the placement of cache atop RAN, and such behaviour is guided by the first term of equations (1) and (9). Finally, the analysis instigated that only the multi-hop cooperation could not guarantee latency reduction. It is necessary to perform a performance analysis of network metrics for the models, as shown in the following.

5.5.2 Network-related metrics

Although the cache hit and total storage values obtained by the multi-hop and network-aware models were higher than the non-cooperative approach, it is evident that only performing the request routing and multi-hop placement do not guarantee the overall QoS. From both MNO and users perspectives, it is not enough just to place caches on HCN BS, and eventually overload a network region, or even worst retrieve a large cache content from the original repository. Given the facts, it is necessary to load balancing the requests, that is, on the one hand, prioritise the consumption of content from caches and, on the other hand, analyse whether consumption through BH is more advantageous depending on the presence of congestion in the RAN.

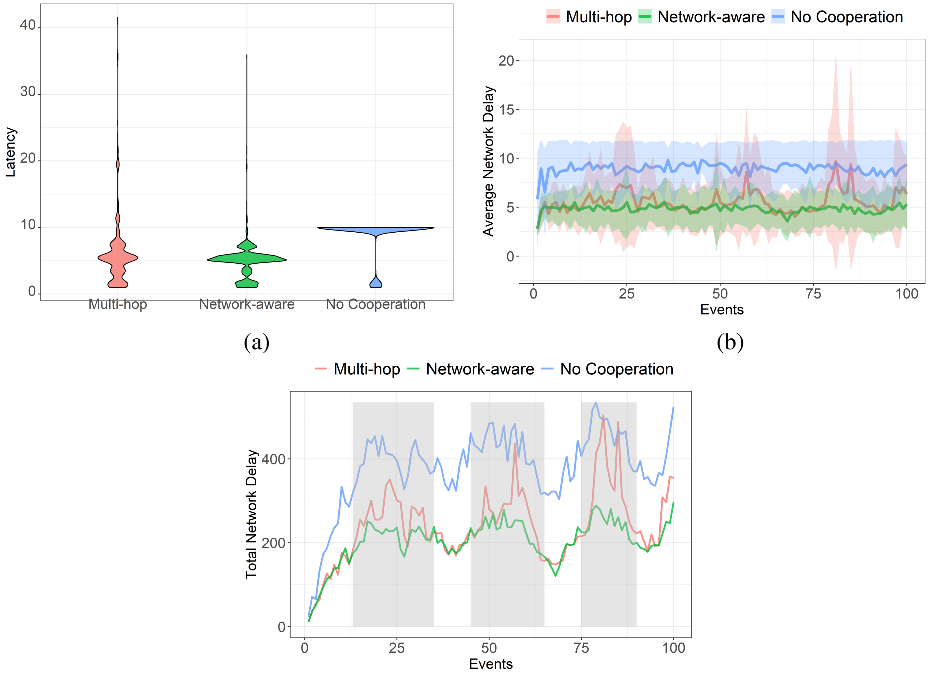
Initially, Figure 4(a) presents the latency distribution using a violin plot. We opt for using a violin plot as the width of each curve corresponds with the approximate frequency of data points in each region (whenever necessary, details and specific values are informed in the text). For the non-cooperative model, most of the requests were carried through BH and, in only 11% of the sample, the latency was less than 10 ms. Such behaviour means that the non-cooperative model has only two choices: consume the content of the BS in which the user is connected with, which in general is a shorter path and consequently with lower latency, or consume the content through the BH, which the latency is approximately 10 times higher. This behaviour corroborates the cache hit results.

The multi-hop and network-aware models have a similar profile as indicated by Figure 4(a). Specifically, for 75% of the multi-hop sample the latency was less than 7 ms, while for the network-oriented model, 75% of the sample was less than 5 ms. Moreover, for the network-oriented model, 99% of the sample was less than 10 ms against 91% for the multi-hop model. In summary, it is observed that the network-oriented model maintained a stable latency.

The multi-hop model inherits the characteristic of positioning as close as possible to the user, behaviour resulting from the first term of equation (9), similar to the first term of equation (1) (objective function of the network-aware model). However, the multi-hop model allocates requests prioritising the content placement and the requests routing closer to the edge of the RAN and, later allocates requests at levels farther from the edge or travels the content through the BH. This choice can group the requests in a single network region, which can overload the network inducing congestion, and consequently increasing the overall latency, as highlighted in Figure 4(b).

Figure 4(c) emphasises the success of the network-aware proposed model to reduce latency in relation to the baseline models considering the total latency per event. It is possible to observe that initially both models have a continuous increase in the total latency due to the first instant in which the requests are allocated, and later there are slopes in the curves. Three main slops deserve emphasis (identified in grey) demonstrating instances where the multi-hop model increases the total latency, sometimes up to the non-cooperative model. Moreover, Figure 4(c) evidences the stability of the latency reduction for the network-oriented model.

Figure 4 Network-related metrics: (a) network latency distribution; (b) average network latency and (c) total network latency (see online version for colours)



(a)

(b)

(c)

Finally, Figure 4(b) corroborates the network-aware performance demonstrating an average latency above 4 milliseconds (specifically, only event 70 obtained a latency greater than the desired threshold) (ITU, 2017). In turn, the same graph evidences the instability of the multi-hop model. The multi-hop curve has peaks that emphasise latency values greater than 4 ms and collaborate to conclude that the non-network-oriented multi-hop model does not prevent the network latency from increasing given the eventual existence of congestion. In turn, the network-oriented model, in general, demonstrates greater latency stability due to the load spreading and balancing effects induced by the objective function. In addition, the network-oriented model searches for links with higher throughput capacities, always looking for lower latency paths, which means that the model avoids aggravating the congestion in the links (as long as possible).

6 Conclusion

The ever-increasing access to multimedia content poses new challenges for mobile network operators, while in parallel, the 5G network standard defines strict latency and high throughput QoS requirements that must be guaranteed for hosting distinct applications. Given the facts, the use of caches closer to mobile devices is a common approach to meet the requirements. Despite all benefits introduced by positioning caches on HCN base stations, this scenario brought a set of challenges to MNO. Specifically, the total storage capacity of BS is a limited and highly requested resource that must be carefully managed, for instance avoiding the creation of unnecessary replicas. In addition, data traffic from multiple applications with distinct QoS requirements are transferred over the HCN fronthaul. This shared and competitive networking scenario poses a barrier to achieve the 5G QoS.

In this context, this work specified and developed a network-oriented cooperative cache policy to reduce latency in HCN. In addition, the cache policy jointly addressed the content placement and request routing challenges. A model was formulated using ILP with storage capacity restrictions and quality-of-service requirements defined by the cache service provider. The model innovates by considering the concurrent background network traffic impact on cache services. Specifically, the model is based on well-known TCP congestion control premises and algorithms to estimate RTT and throughput values on competitive and decentralised scenarios. Numerical simulations demonstrated the benefits introduced by the network-aware model. For performing a fair comparison, the baseline models (which represents the specialised literature) were improved to identify the HCN links load dynamically. Finally, the network-oriented and cooperative cache policy proved to be efficient according to the analyses carried out.

The promising results obtained indicate some future directions. Initially, the model can be extended to deal with other QoS requirements commonly presented in the edge of the Internet (e.g., processing capacity, delay-sensitive applications), while a second line indicates an actual and practical implementation on testbeds. Finally, solving an ILP is known to be *NP-hard*, and although the proposed cache policy demonstrated a more stable latency and efficient resource usage when compared to counterparts, it imposes a scalability barrier. The ILP-based cache policy paved the road and demonstrated how we can achieve efficient network-aware cache provisioning and request routing in HCN, however the effective use on large-scale scenarios require the application of techniques to relax the constraints, as well as the proposition of approximation heuristics to eventually reduce the search space.

Acknowledgements

This work was supported by the National Council for Scientific and Technological Development (CNPq) grant 311245/2021-8, the Santa Catarina State Research and Innovation Support Foundation (FAPESC), Santa Catarina State University (UDESC), and developed at Laboratory of Parallel and Distributed Processing (LabP2D).

References

Agiwal, M., Roy, A. and Saxena, N. (2016) ‘Next generation 5g wireless networks: a comprehensive survey’, *IEEE Communications Surveys and Tutorials*, Vol. 18, No. 3, pp.1617–1655.

Alves, M. and Koslovski, G.P. (2022) 'Joint cache placement and request routing optimization in heterogeneous cellular networks', *2022 IEEE Symposium on Computers and Communications (ISCC)*, Rhodes, Greece, pp.1–6.

Andrews, J.G. (2013) 'Seven ways that hetnets are a cellular paradigm shift', *IEEE Communications Magazine*, Vol. 51, No. 3, pp.136–144.

Ayenew, T.M., Xenakis, D., Passas, N. and Merakos, L. (2021) 'Cooperative content caching in mec-enabled heterogeneous cellular networks', *IEEE Access*, Vol. 9, pp.98883–98903.

Bastug, E., Bennis, M. and Debbah, M. (2014) 'Living on the edge: the role of proactive caching in 5g wireless networks', *IEEE Communications Magazine*, Vol. 52, No. 8, pp.82–89.

Brakmo, L.S. and Peterson, L.L. (1995) 'TCP vegas: End to end congestion avoidance on a global internet', *IEEE Journal on Selected Areas in Communications*, Vol. 13, No. 8, pp.1465–1480.

Breslau, L., Cao, P., Fan, L., Phillips, G. and Shenker, S. (1999) 'Web caching and zipf-like distributions: evidence and implications', *IEEE INFOCOM '99. Conference on Computer Communications. Proceedings. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. The Future is Now (Cat. No.99CH36320)*, Vol. 1, pp.126–134.

Cardwell, N., Cheng, Y., Gunn, C.S., Yeganeh, S.H. and Jacobson, V. (2016) 'Bbr: Congestion-based congestion control', *ACM Queue*, Vol. 14, September–October, pp.20–53.

Chiu, D-M. and Jain, R. (1989) 'Analysis of the increase and decrease algorithms for congestion avoidance in computer networks', *Computer Networks and ISDN Systems*, Vol. 17, No. 1, pp.1–14.

Chowdhury, M., Rahman, M.R. and Boutaba, R. (2012) 'Vineyard: Virtual network embedding algorithms with coordinated node and link mapping', *IEEE/ACM Transactions on Networking*, Vol. 20, No. 1, pp.206–219.

Damnjanovic, A., Montojo, J., Wei, Y., Ji, T., Luo, T., Vajapeyam, M., Yoo, T., Song, O. and Malladi, D. (2011) 'A survey on 3gpp heterogeneous networks', *IEEE Wireless Communications*, Vol. 18, No. 3, pp.10–21.

De Souza, F.R., Miers, C.C., Fiorese, A. and Koslovski, G.P. (2017) 'QoS-aware virtual infrastructures allocation on sdn-based clouds', *2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, Madrid, Spain, pp.120–129.

de Souza, F.R., Miers, C.C., Fiorese, A., de Assunção, M.D. and Koslovski, G.P. (2019) 'QVIA-SDN: towards qos-aware virtual infrastructure allocation on SDN-based clouds', *J. Grid Computing*, Vol. 17, pp.447–472, https://doi.org/10.1007/s10723-019-09479-x

Dehghan, M., Jiang, B., Seetharam, A., He, T., Salonidis, T., Kurose, J., Towsley, D. and Sitaraman, R. (2017) 'On the complexity of optimal request routing and content caching in heterogeneous cache networks', *IEEE/ACM Transactions on Networking*, Vol. 25, No. 3, pp.1635–1648.

Harutyunyan, D., Bradai, A. and Riggio, R. (2018) 'Trade-offs in cache-enabled mobile networks', *2018 14th International Conference on Network and Service Management (CNSM)*, Rome, Italy, pp.116–124.

International Telecommunication Union (ITU) (2017) *Minimum Requirements Related to Technical Performance for IMT-2020 Radio Interface(s)*, Technical Report, International Telecommunication Union (ITU).

Jiang, W., Feng, G. and Qin, S. (2017), 'Optimal cooperative content caching and delivery policy for heterogeneous cellular networks', *IEEE Transactions on Mobile Computing*, Vol. 16, No. 5, pp.1382–1393.

Kabir, A., Rehman, G., Gilani, S.M., Kitindi, E.J., Ul Abidin Jaffri, Z. and Abbasi, K.M. (2020) 'The role of caching in next generation cellular networks: A survey and research outlook', *Trans. Emerg. Telecommun. Technol.*, Vol. 31, No. 2, pp.1–25.

Kamel, M., Hamouda, W. and Youssef, A. (2016) 'Ultra-dense networks: a survey', *IEEE Communications Surveys Tutorials*, Vol. 18, No. 4, pp.2522–2545.

Karp, R.M. (1975) ‘On the computational complexity of combinatorial problems’, *Networks*, Vol. 5, No. 1, pp.45–68.

Khreishah, A., Chakareski, J. and Gharaibeh, A. (2016) ‘Joint caching, routing, and channel assignment for collaborative small-cell cellular networks’, *IEEE Journal on Selected Areas in Communications*, Vol. 34, No. 8, pp.2275–2284.

Li, H., Li, X., Sun, C., Fang, F., Fan, Q., Wang, X. and Leung, V. C.M. (2024) ‘Intelligent content caching and user association in mobile edge computing networks for smart cities’, *IEEE Transactions on Network Science and Engineering*, Vol. 11, No. 1, pp.994–1007.

Li, X., Wang, X., Li, K., Han, Z. and Leung, V.C.M. (2017) ‘Collaborative multi-tier caching in heterogeneous networks: Modeling, analysis, and design’, *IEEE Transactions on Wireless Communications*, Vol. 16, No. 10, pp.6926–6939.

Liu, Y., Liu, J., Argyriou, A., Wang, L. and Xu, Z. (2021), ‘Rendering-aware VR video caching over multi-cell mec networks’, *IEEE Transactions on Vehicular Technology*, Vol. 70, No. 3, pp.2728–2742.

Lyu, X., Ren, C., Ni, W., Tian, H., Liu, R.P. and Tao, X. (2021) ‘Distributed online learning of cooperative caching in edge cloud’, *IEEE Transactions on Mobile Computing*, Vol. 20, No. 8, pp.2550–2562.

Ordonez-Lucena, J., Ameigeiras, P., Lopez, D., Ramos-Munoz, J.J., Lorca, J. and Folgueira, J. (2017) ‘Network slicing for 5g with sdn/nfv: Concepts, architectures, and challenges’, *IEEE Communications Magazine*, Vol. 55, No. 5, pp.80–87.

Paschos, G.S., Iosifidis, G., Tao, M., Towsley, D. and Caire, G. (2018) ‘The role of caching in future communication systems and networks’, *IEEE Journal on Selected Areas in Communications*, Vol. 36, No. 6, pp.1111–1125.

Pham, Q-V., Fang, F., Ha, V.N., Piran, M.J., Le, M., Le, L.B., Hwang, W-J. and Ding, Z. (2020) ‘A survey of multi-access edge computing in 5g and beyond: Fundamentals, technology integration, and state-of-the-art’, *IEEE Access*, Vol. 8, pp.116974–117017.

Pu, L., Jiao, L., Chen, X., Wang, L., Xie, Q. and Xu, J. (2018) ‘Online resource allocation, content placement and request routing for cost-efficient edge caching in cloud radio access networks’, *IEEE Journal on Selected Areas in Communications*, Vol. 36, No. 8, pp.1751–1767.

Rafique, W., Hafid, A.S. and Cherkaoui, S. (2023) ‘Softcaching: a framework for caching node selection and routing in software-defined information centric internet of things’, *Computer Networks*, Vol. 235, pp.109966.

Shanmugam, K., Golrezaei, N., Dimakis, A.G., Molisch, A.F. and Caire, G. (2013) ‘Femtocaching: Wireless content delivery through distributed caching helpers’, *IEEE Transactions on Information Theory*, Vol. 59, No. 12, pp.8402–8413.

Sheng, M., Xu, C., Liu, J., Song, J., Ma, X. and Li, J. (2016) ‘Enhancement for content delivery with proximity communications in caching enabled wireless networks: architecture and challenges’, *IEEE Communications Magazine*, Vol. 54, No. 8, pp.70–76.

Somesula, M.K., Mothku, S.K. and Annadanam, S.C. (2023) ‘Cooperative service placement and request routing in mobile edge networks for latency-sensitive applications’, *IEEE Systems Journal*, Vol. 17, No. 3, pp.4050–4061.

Song, Y., Wo, T., Yang, R., Shen, Q. and Xu, J. (2021) ‘Joint optimization of cache placement and request routing in unreliable networks’, *Journal of Parallel and Distributed Computing*, Vol. 157, pp.168–178.

Tian, Y., Xu, K. and Ansari, N. (2005) ‘Tcp in wireless environments: problems and solutions’, *IEEE Communications Magazine*, Vol. 43, No. 3, pp.S27–S32.

Wang, S., Zhang, X., Zhang, Y., Wang, L., Yang, J. and Wang, W. (2017) ‘A survey on mobile edge networks: Convergence of computing, caching and communications’, *IEEE Access*, Vol. 5, pp.6757–6779.

Wu, H., Fan, Y., Wang, Y., Ma, H. and Xing, L. (2021) ‘A comprehensive review on edge caching from the perspective of total process: Placement, policy and delivery’, *Sensors*, Vol. 21, No. 15, p.5033.

Xie, T., Thakkar, S., He, T., McDaniel, P. and Burke, Q. (2023) ‘Joint caching and routing in cache networks with arbitrary topology’, *IEEE Transactions on Parallel and Distributed Systems*, Vol. 34, No. 8, August, pp.2237–2250.

Notes

[1] An early version of the cache policy and preliminary results (calibration and proof-of-concept) were published at IEEE Symposium on Computers and Communications (ISCC) 2022 (Alves and Koslovski, 2022). Details are presented in Section 2.3.

[2] Available at https://www.gurobi.com