Optimal Content Management and Dimensioning in Wireless Networks

Jonatan Krolikowski

université PARIS-SACLAY

NNT : 2018SACLS452

Thèse de doctorat

UNIVERSITÉ PARIS SUD
Comprendre le monde, construire l'avenir

Optimal Content Management and Dimensioning in Wireless Networks

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université Paris-Sud au sein du Laboratoire des Signaux et Systèmes

Ecole doctorale n°580 Sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat : Réseaux, Information et Communications

Thèse présentée et soutenue à Paris, le 06/11/2018, par

JONATAN KROLIKOWSKI

Composition du Jury :

Isabelle GUERIN LASSOUS Professeur, Université Claude Bernard Lyon 1 (Département informatique)	Présidente, Rapporteur
André-Luc BEYLOT Professeur, Institut de Recherche en Informatique de Toulouse (ENSEEIHT-IRT)	Rapporteur
Jocelyne ELIAS Maître de Conférences, Université Paris Descartes (LIPADE)	Examinatrice
Dario ROSSI Professeur, Télécom ParisTech (INFRES)	Examinateur
Salah Eddine ELAYOUBI Maître de conférences, CentraleSupélec (L2S)	Examinateur
Georgios IOSIFIDIS Assistant Professor, The University of Dublin (School of Computer Science and Statistics)	Examinateur
Marco DI RENZO Chargé de Recherche, CentraleSupélec (L2S)	Directeur de Thèse
Anastasios GIOVANIDIS Chargé de Recherche, Sorbonne Université (LIP6)	Co-encadrant de Thèse

Optimal Content Management and Dimensioning in Wireless Networks

Jonatan Krolikowski

Abstract

The massive increase in cellular traffic poses serious challenges to all actors concerned with wireless content delivery. While network densification provides access to additional users, high-speed and high-capacity backhaul connections are expensive. Caching popular content at the network edge promises to offload user traffic from these congestion prone connections as well as from the data centers in the backbone network.

This thesis proposes a business model in which a mobile network operator (MNO) pre-installs and maintains caches at its wireless equipment (Cache-equipped Base Stations, CBSs). Memory space together with computational capabilities is then leased to content providers (CPs) that want to bring their content closer to the user. For a financial compensation, a CP can then offload traffic from its data center and improve user Quality of Service. The CP makes content placement decisions based on user traffic and content popularity data. In the delivery phase, users can be served from the caches in case they are associated to stations that have the requested content cached.

This work investigates three aspects of the proposed business model: The first research question focuses on user association as a central element to the edge caching scheme. Cache-aware user association policies can allow for users in coverage overlap areas to be associated to a CBS that holds the requested content rather than conventionally to the one that provides the strongest signal. The thesis proposes an original decentralized algorithm for user association called Generalized Bucket-filling that allows gains beyond

maximizing the hit ratio. Performance metrics such as network throughput and load balancing of users among CBSs are taken into account. Experiments show that cache-aware user association a) increases the hit ratio b) without overloading single CBSs while c) providing high system throughput.

The second problem treated considers a single CP that needs to decide how much cache space to lease at each CBS for a fixed price, and what content to place. Its choices should be based on estimates of file popularity as well as MNO user association policy. The cache leasing and content placement problem is formulated as a non-linear mixed-integer problem (NLMIP). In its solution, the problem is separated into a linear discrete CP subproblem and a nonlinear continuous subproblem using Benders decomposition. The CP and the MNO cooperate, helping the CP to make optimal decisions that benefit both parties: The CP maximizes its savings from caching while the MNO can find the optimal cache price and receive the maximum financial compensation.

A third research question widens the focus to the interaction between several CPs and one MNO. Now, the MNO does not set a fixed price per memory unit but instead reacts to CP demands for memory space that depend on the savings they can achieve from caching.

Résumé en Français

Introduction

Le Cisco Visual Networking Index prévoit que le trafic Internet mobile mondial sera multiplié par sept entre 2016 et 2021. Il devrait atteindre 48,3 EB par mois, passant de 7 à 17% du trafic IP total au cours de cette période. Le trafic vidéo Internet mobile devrait même être multiplié par 9, ce qui représenterait 21% du trafic vidéo total. Comment faire face à l'immense défi posé par cette forte augmentation de demande ?

Cette thèse explore caching (ou : mise en cache) à la bordure des réseaux cellulaires. Le contenu populaire est temporairement stocké sur les nœuds d'accès sans fil de manière distribuée afin de rapprocher le contenu à l'utilisateur. La mise en cache est conditionnée par divers aspects techniques et économiques des réseaux sans fil actuels et futurs, posant ainsi des défis distincts aux systèmes de mise en cache traditionnels des réseaux fixes. Au cœur de cette thèse, un modèle économique est proposé qui prend en compte ces conditions particulières pour servir les intérêts de toutes les parties prenantes de la diffusion de contenu. Caching est un concept bien connu. Il semble naturel d'étendre le concept de stockage de contenu distribué aux réseaux sans fil. Cependant, la diffusion par les réseaux sans fil présente un ensemble de défis particulier dans le contexte de caching. Ces défis, et comment les surmonter, sont le sujet de cette thèse.

Portée de la thèse et contributions

Dans cette thèse, trois questions fondamentales de la mise en cache à la bordure du réseau sans fil sont abordées : dimensionnement et économie du cache, placement du contenu, et association des utilisateurs. L'accent est mis sur les aspects économiques de la mise en cache à la bordure du réseau cellulaire sans fil. Les caches aux BSs (stations de base, anglais: base stations) sont installés par le MNO (opérateur de réseau mobile, anglais: mobile network operator). Les BSs sont équipées (ou non) avec des liaisons de transport à grande vitesse et peuvent avoir des portées différentes, mais nous supposons qu'elles utilisent des technologies d'accès radio équivalentes. Le MNO propose aux CPs l'espace de cache disponible en échange de compensation financière (leasing à prix fixe ou à prix adaptatif). Les CPs sont supposés prendre des décisions rationnelles en ce qui concerne l'espace de cache acquis ainsi que le placement de contenu. Des fichiers entiers sont placés dans une phase de pré-extraction (pendant les heures creuses), car c'est possible qu'il n y a pas de connexions de transport à grande vitesse. Pour les décisions de placement raisonnables, nous supposons que nous disponons des données sour la densité du trafic et de la popularité du contenu. L'importance des stratégies d'association d'utilisateurs de l'MNO, qu'elles soient ou non réactives à caching, est un élément central des décisions du CP. L'hypothèse sous-jacente est que les utilisateurs dans des zones de chevauchement de couverture peuvent potentiellement être associés à l'une des stations de couverture, en tenant compte du compromis entre la disponibilité du contenu local et les conditions radio.

En particulier, les contributions de cette thèse sont:

- La thèse propose et analyse un modèle économique pour la location de caches à la bordure du réseau aux CPs. De cette façon, on considère les incitations économiques liées à la mise en cache des contours qui

conditionnent la viabilité du concept.

- On étudie les politiques d'association des utilisateurs prenant en compte le cache. Lorsque des utilisateurs situés dans des zones de chevauchement de couverture peuvent être associés à n'importe quel CBS (stations de base avec cache, anglais: cache-equipped base station) couvrant, ils peuvent potentiellement accéder au contenu de plusieurs caches. La thèse présente de nouvelles stratégies d'association d'utilisateurs prenant en compte le cache, qui permettent un compromis entre la force du signal et la disponibilité du contenu mis en cache. Les stratégies peuvent atteindre différents objectifs tels que l'équilibrage de la charge des CBSs ou la maximisation du débit. La thèse démontre que ces stratégies améliorent les performances de la mise en cache à la bordure du réseau sans fil.
- Fondée sur une association d'utilisateurs prenant en compte le cache, cette thèse, à la connaissance de l'auteur, est la première à optimiser conjointement les décisions de location, d'emplacement de contenu et d'association d'utilisateurs sur la base de données de trafic et de popularité de contenu précises, en supposant que des fichiers entiers soient mis en cache. La politique de tarification optimale est déterminée, ce qui rend l'activité de mise en cache plus rentable pour le MNO.
- L'influence de la concurrence entre plusieurs CPs sur la tarification du cache est explorée.

Résumé des Résultats

Dans une première étape en chapitre 4, l'association des utilisateurs prenant en compte le cache est examinée. Un algorithme efficace appelé Generalized Bucket-filling est développé pour le calcul de l'association des utilisateurs prenant en compte le cache optimisant différents critères de performance.

Les principales conclusions de l'application de l'algorithme sont les suivantes:

- L'association es utilisateurs prenant en compte le cache offre des grands avantages par rapport à l'association des utilisateurs classique pour un placement de contenu donné. En particulier, les mesures de performance telles que le taux de réussite ou le débit du réseau sont considérablement améliorés.
- Les performances des associations des utilisateurs sont améliorées, en particulier si les CBS voisins stockent un contenu différent.
- Lorsque l'objectif est l'équilibrage de la charge des utilisateurs parmi les CBS, il est possible d'atteindre un taux de réussite élevé sans surcharger les CBSs, en transférant le trafic supplémentaire vers les stations moins utilisées.
- Les meilleures performances par rapport à l'association des utilisateurs sont obtenues lorsque le contenu est placé d'une manière qui anticipe l'association des utilisateurs.
- L'association es utilisateurs prenant en compte le cache fonctionne mieux lorsque les stations voisines utilisent différentes bandes de fréquences, ce qui réduit l'influence de l'interférence.

Dans une deuxième étape, chapitre 5, un scénario est examiné dans lequel un MNO loue de la mémoire cache à un CP. Le problème de la location optimale de cache et du placement de contenu utilisant une association des utilisateurs prenant en compte le cache est décomposé en une partie MNO et une partie CP. La stratégie de location et de placement qui en résulte, qui optimise le CLCP (Cache Leasing and Content Placement problem) lorsque l'association des utilisateurs Opt- h est appliquée, est également appelée Opt- h. Des expériences montrent la supériorité de Opt- h par rapport aux décisions optimales de location de cache et de placement de contenu utilisants la stratégie d'association cache-inconsciente Closest. Les conclusions sont:

- Le plus grand soient les zones de chevauchement de couverture des

CBSs, est le plus élevé soit le prix de location du cache, le meilleur est la performance de Opt- h par rapport à Closest.

- La location de cache et le placement de contenu suivant Opt- h fournit un catalogue de contenu mis en cache plus varié que Closest.
- Le MNO peut identifier le point opérationnel de la tarification qui maximise ses revenus. Le prix optimal dépend du rayon de couverture des CBSs.
- L'efficacité de la mise en cache à la bordure du réseau dépend des statistiques sur la popularité du contenu. Le plus les statistiques sont faussées vers moins de fichiers populaires, le plus la mise en cache est efficace.
- Alors que la stratégie de mise en cache proposé dépend des statistiques sur la popularité du contenu, des petites inexactitudes statistiques n'ont que des effets mineurs sur l'efficacité de la stratégie.
- La location et le placement basés sur Opt- h sont supérieurs à Closest,en particulier dans les scénarios dans lesquels l'interférences inter-cellule est réduite grâce à l'utilisation de bandes de fréquences orthogonales en stations voisines.

Au-delà des résultats de Chapitres 4 et **??**, Chapitre 6 présente un moyen d'évaluer des scénarios dans lesquels plusieurs CPs disputent les ressources de cache d'un opérateur de réseau. Dans un tel scénario, les CPs peuvent agir en tant que preneurs de prix (price takers), en acceptant le prix de leasing indiqué par le MNO, ou ils peuvent anticiper le prix. Dans ce dernier cas, les CPs préemptent les réactions des CPs concurrents, supposant que l'opérateur MNO souhaite louer tout l'espace de cache disponible. Dans le cas où les CP anticipent les prix, tout équilibre du jeu résultant ne perd pas plus de 25% d'efficacité par rapport à l'optimum social.

Acknowledgements

I would like to thank

- my PhD co-advisor Dr. Anastasios Giovanidis for the ideas the discussions and for making all this possible,
- my PhD advisor Dr. Marco Di Renzo for his helpful comments and his patience,
- all members of my jury for the interesting questions,
- Prof. Dr. Alain Sibille for his calm helpfulness,
- Digiteo/Digicosme for giving me the opportunity,
- Télécom ParisTech for giving me an office and a great work environment,
- friends and colleagues at Télécom ParisTech/Lincs for all the coffees and the interesting conversations,
- my parents for all the support,
- S. for joy and love.

It was a great time!

Contents

Chapter 1

Introduction

The Cisco Visual Networking Index [Cis17] predicts that global mobile internet traffic will increase by factor seven between 2016 and 2021. It is expected to reach 48.3 EB per month, rising from 7 to 17 percent of total IP traffic in said period. Mobile consumer internet video traffic is even expected to grow by factor 9, leading to a 21 percent share of all video traffic. How can the immense challenge posed by this massively increased demand be met?

This thesis explores *wireless edge caching* in cellular networks. *Popular content* is *temporarily* stored at *wireless access nodes* in a *distributed* fashion to bring the content closer to the user. Wireless edge caching is conditioned by various technical and economical aspects of current and future wireless networks, thus posing challenges distinct to traditional caching schemes in fixed networks. At the core of the thesis, a business model is proposed that takes these particular conditions into account to serve the interest of all stakeholders in wireless content delivery.

Caching is a well-known concept. It seems natural to extend the concept of distrubuted content storage into wireless networks. However, wireless delivery presents a particular set of challenges to caching. These challenges, and how to overcome them, are the subject-matter of this thesis.

1.1 Current and Future Wireless Networks

Backhaul connection of wireless access points is a bottleneck in content delivery, particularly with the introduction of heterogeneous networks or *HetNets* [DRGC13]. However, interference in the wireless channel is also a problem that comes from network densification. Spectral efficiency can be significantly enhanced with increased spatial reuse. This, however, comes at the cost of increased interference. On the other hand, interference can be reduced by avoiding the simultaneous use of frequencies in neighboring BSs when a user is located at the cell edge, i.e. potentially receiving the signal and the interference with similar channel gain.

Cloud radio access networks (C-RAN) are part of the future 5G architecture. In C-RAN, remote radio heads (RRHs) are controlled and coordinated by baseband units (BBUs) [NLPW14, YCXW15]. In this context, the backhaul bottleneck is overcome via highspeed fiber connections between RRHs and BBUs. Placing caches at the BBUs allows cached content to be served no matter which subset of RRHs is actively participating in the transmission. C-RAN promises to cut certain costs compared to similar coverage by MBSs:

- Cooperative transmission schemes allow for the reduction of interference.
- Centralization of computational capacities are more energy-efficient than distributed MBSs that each need their independent infrastructure.

1.2 Caching at the Edge

To introduce caches into wireless networks, the location of these caches within the networks needs to be decided. Consider a network transmitting content from a central content storage to a wireless user as can be seen in Figure 1.1. The centralized storage is denoted by DC for data center. The content is

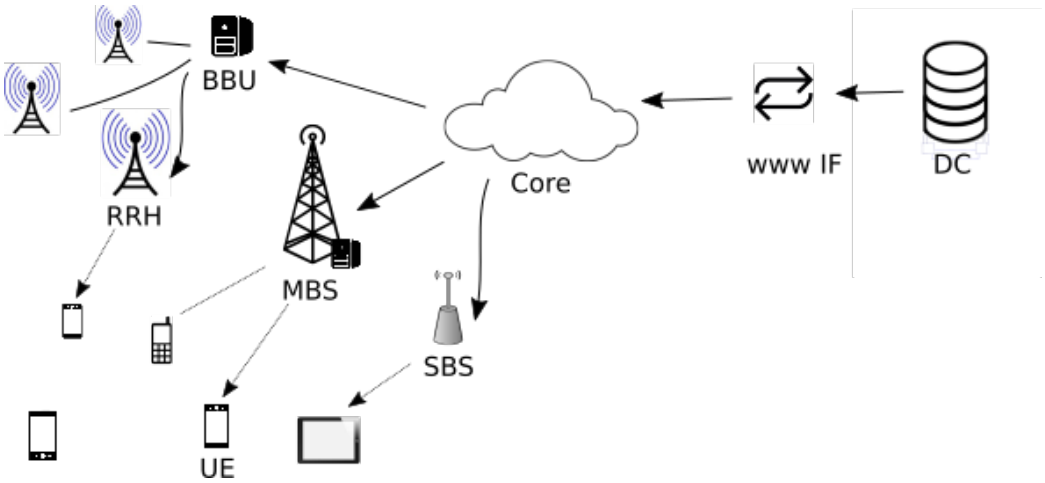

Figure 1.1: Content delivery in a wireless network.

delivered via an interface (www IF) to the wireless core network. From there, in a heterogeneous network, content can be delivered to the user equipment (UE) through different wireless access nodes: macro BSs (MBSs), small BSs (SBSs) or C-RAN (BBU and several RRHs).

Where should caches be installed? The following three possibilities are considered in the literature:

Core Network or Baseband Unit

Caching in the LTE core network using TCP redundancy elimination has been proposed in [WJP+13]. The advantage of caching deeper in the network is that users associated to several BSs can be routed to the cache to make use of the cached content. However, it also implies that the caches are located behind the backhaul connection. Thus, such caches are not able to alleviate the expected bottleneck link of wireless networks. When caches are installed at the baseband unit in C-RAN architecture, they are closer to the user but as a consequence potentially accessible for fewer users.

Mobile Devices

Caching at mobile devices has been proposed in [GMDC13] making use of device-to-device (D2D) communication. Caching on mobile devices has the advantage of potential individualized caching strategies. D2D content transmission futhermore allows wider use of these caches without use of the BS backhaul connections. Still, it needs to be taken into account that device caches need to be comparatively small, and D2D file transfer depends on the relatively weak signal emitted by mobile user devices and is exposed to mobility of both the sending and the receiving device.

Base Stations

The original work suggesting to cache at the BSs is *FemtoCaching* by Shanmugam et al [SGD+13], i.e. *edge caching*. This work proposes to install wireless distributed caching helpers. These helpers are wireless access points equipped with large cache memory but have no high-speed backhaul connection. The main advantage of edge caching is that the fixed BS infrastructure allows for the installation of large inexpensive caching infrastructure that can serve large numbers of users and, in case the user-requested content is cached, offloads traffic from the scarce backhaul resources. The drawbacks of edge caching are that (1) traffic in the coverage area of BSs is difficult to predict and (2) conventional user association does not take caching decisions into account.

Note that the installation of caches at BSs changes their functionality. Conventionally, they are able to measure wireless channel conditions as well as to initiate, accomodate and manage communication requests. When equipped with caches, a controller is needed that can identify the requested content and determine whether it is cached. This controller can either be placed at the BS itself or deeper in the network, potentially even at the CP's content storage. When the controller induces content delivery from the cache,

a mechanism is necessary for the packetization as well as for the encryption of the content.

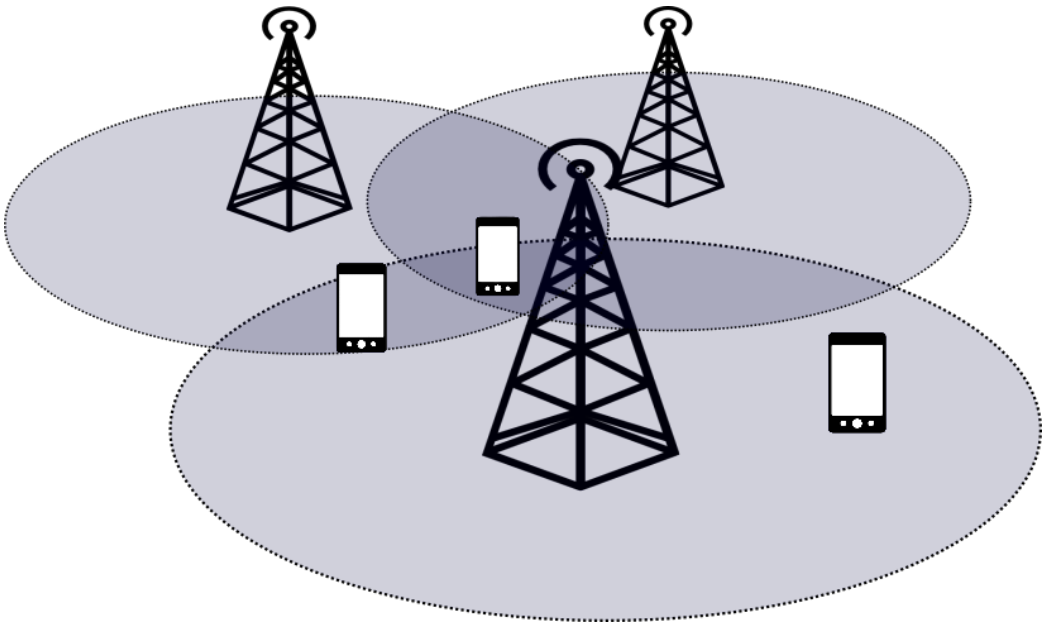

Figure 1.2: Three BSs whose coverage areas overlap. Users in the overlap regions can potentially be associated to any covering BS.

1.3 Edge Caching Benefits

This work proposes the installation of caches at base stations (or, in the following equivalently: wireless access nodes), turning them into Cache-equipped base stations (CBSs). The main reasons for this decision are:

- **Load reduction** on upstream infrastructure: Cache hits at CBSs reduce the load on upstream infrastructure. This includes the backhaul connections that are identified as probable bottlenecks in content delivery but also the CPs' content storage entities. Redundant transmissions can be avoided on long network paths: over time, multiple requests for the same content can be served from the cache while the delivery from the source to the cache only needs to occur once.

- **Reliability**: In case of overload or other system failure at the central content storage, some system performance can be maintained. At the same time, base stations provide more reliable data transmissions than D2D connections in case of caching on user equipment.
- **Delay reduction**: The closer to the user content is cached, the lower is the delivery delay, thus improving QoS. Especially when BSs are not equipped with high-speed backhaul connections, local caches provide significant delay reduction. While caching on user equipment reduces the physical distance even further, the delivery delay via D2D connections is dependent not only on the position and mobility of the receiving user but also on the sender.
- **Multi-coverage benefits**: Particularly in dense urban networks, coverage areas of CBSs overlap. Given a cache-aware user association policy, users in multi-coverage areas (see Figure 1.2) potentially have access to all covering caches. Then, a diversified content in neighboring stations is particularly useful.
- **Localization**: Caches at the wireless edge serve only a relatively small area, thus traffic is spatially fine grained. The caches can thus adapt better to local demands. While device caching is ultimately more fine grained, CBSs allow the use of content to several users at the same location.
- **Adaptation of Video Quality**: Videos are usually offered in different image qualities that are provided in different chunks of files. The quality of the delivered video is adapted to the user channel quality and congestion. By storing a video in different quality levels in a cache, a user with good channel conditions can enjoy high-quality video without the need of high-speed download through the backhaul. When the channel quality decreases, the video quality adapts.

1.4 Edge Caching Challenges

There are several challenges that need to be faced to put edge caching into practice.

- **Cache Dimensioning and Economics**: Caches need to be installed, physically maintained and supplied with energy. The example of Netflix Open Connect Appliances (OCA, see Section A.3.2) show an example in which a CP maintains caching infrastructure. At BSs, however, physical space is limited and generic boxes are not as easily placed. This suggests that edge caches should be installed and managed by MNOs. The optimal dimensions of edge caches are then to be determined by the MNOs. Bigger caches can contain more content and thus increase the hit ratio, but also incur higher financial and energy costs. While it is possible to operate edge caches in a content-agnostic way, e.g. by applying Last Recently Used (LRU) as a replacement policy, pervasive encryption of communication between clients and CPs make the involvement of CPs necessary. Thus, an MNO maintaining caches needs to decide which CPs should participate in the cache operation, and to which price. In short, a business model for edge caching in which MNOs offer cache space to CPs needs to be developped.
- **Content Placement**: The usefulness of caches is maximized when they store content that will likely be downloaded. Thus, the question what content should be placed in the caches is central to the effectiveness of caching.

  There are two different notions how to approach this problem: In the online caching model, when a cache miss occurs, the request is forwarded to the central storage. When the requested file passes through the network node while transmitted to the user, it is cached (or not). The policy by which it is decided if the file is cached and which file is

evicted from the cache instead is called *replacement policy*. The most commonly known policy replaces the Least Recently Used (LRU) file.
Prefetching means placing content into caches in a placement phase that is planned to occur before the content delivery. The basis for the placement decisions is expected traffic as well as content popularity. The caches are filled in off peak hours, and the cached content is not changed during peak hours due to a lack of highspeed backhaul connections. The simplest placement policy is to cache the globally or locally most popular content. However, the opportunity to potentially access more than one cache allows for more involved content placement policies.
Finding an efficient content placement/replacement scheme is essential for the viability of edge caching.

- **User Association**: For cache hits to occur, user requests need to be routed to caches that hold the content. For users that can only associate to one CBS, conventionally the one with the strongest signal, the effect of caching depends only on the content placement decisions at that particular cache. Particularly in dense heterogeneous networks, however, several wireless access nodes provide sufficient signal strength. Thus, users potentially can be associated to any such node. Then, a *cache-aware* user association policy can associate users to CBSs that cache the requested content, trading off signal strength for cache hit. Another aspect that user association policies can take into account is load-balancing since the radio resources of access nodes are limited. Proposed user association schemes such as coordinated multi-point help realize such user association.
  In conventional web caching, users can be routed to caches using TCP/IP. In contrast, wireless edge caching requires users to be associated to a particular node to access its cache. This is a process that precedes the

establishment of a TCP/IP connection. Thus, novel cache-aware user association policies need to be explored to bring edge caching to its full potential.

- **Traffic and Popularity Measurement**: A good prediction of user traffic is basic for caching business decisions: The more users are likely to access a cache, the greater the variety of content requests, thus rendering a bigger investment in the cache more reasonable. Furthermore, in case content is placed in a prefetching phase, popularity data is required to make optimal placement decisions. Content popularity varies both in time and space. Machine-learning problems such as user type classification and demand prediction are promising related subjects: They are already applied to learn the preferences of individual users. This can be extended to spatio-temporal content popularity in wireless networks.
- **Mobility**: Users are typically not static in wireless networks. This implies that radio conditions change, and handoffs between radio access points occur. The impact of these handoffs on caching decisions need to be explored. Furthermore, user mobility is of particular interest in device caching schemes. One promising research direction are ad-hoc networks of cache-equipped vehicles.
- **Fractional Caching**: In conventional caching schemes, entire files are cached or the files might be partitioned into smaller chunks. Netflix and Youtube partition their videos into partial files of down to 2 seconds each. In this sense, the splitting of content into various files is already common practice [MSR+17]. Caching, for example, the first segments of videos can be a reasonable strategy when users regularly do not watch videos in their entirety. Following this logic, the total transmitted load can be reduced by partitioning content and delivering only the parts that are actually required.

In these cases of *uncoded* caching, the non-cached content entirely needs to be retrieved from upstream. Broadcast opportunities can be exploited when several requests for the same content coincide. With *coded* caching schemes, broadcasting is advantageous even if users simultaneously request different content. This can be achieved through coordinated prefetching of partial files and computational effort retrieving information from coded information transmission.

- **Privacy and Verification**: Since May 2018, the General Data Protection Regulation (GDPR) is in effect in the European Union (EU). It requires strong measures to protect private customers's personal data. It is built on "principles of data protection by design and data protection by default" [Eur16]. At least for the EU market, this implies that the user behaviour in relation to a CP should not be known to any third party. In parcatice, a large share of data connections are end-to-end (E2E) encrypted. The challenge for caching schemes is to enable such secure and private connections while retaining the benefits that arise from caching.

  Another aspect of private connections between user and CP is the verification of user identity for the distribution of restricted content. Especially subscription based video and audio distribution platforms such as Netflix and Spotify need to make sure that only their clients have access to their content.
- **Net neutrality**: Net neutrality "prohibits Internet service providers from speeding up, slowing down or blocking Internet traffic based on its source, ownership or destination" [KWW13]. While the United States administration has recently distanced itself from the idea of net neutrality, EU law implements some aspects of it [Par15]. There is a legal debate if MNOs can be seen as regular Internet Service Providers (ISPs) and thus, if the net neutrality principle is applicable to wireless

content delivery [Yoo17]. While a detailed legal analysis of the viability of the proposed business model needs to be performed before its implementation, such deliberations go beyond the scope of this thesis.

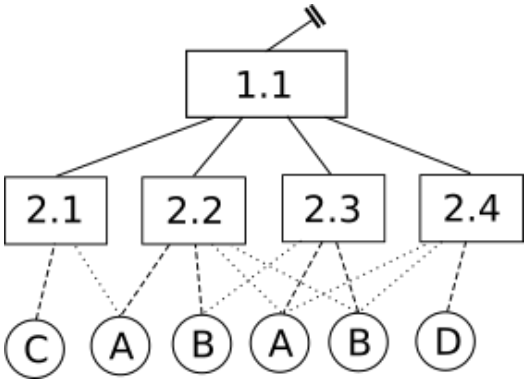

Figure 1.3: A tree-like network

1.5 An Example

The complex relationship between the three aspects installation of caches, content placement and user association is illustrated by the simplified network shown in Figure 1.3. From the bottom, several CP users request content denoted by letters A to D. The dashed lines connect to the respective BS with the strongest signal. Some users are located in coverage overlap areas. In these cases, the dotted lines connect to the BSs that provide a weaker but sufficient signal for the users to associate to. The BSs 2.1 to 2.4 are connected via backhaul links to the inner network node 1.1. Node 1.1 has a connection to the backhaul data center in which files A to D are stored. Each hop that content traverses adds to the download delay.

Assume first that a large cache is installed at node 1.1. Then, if all content A to D is stored there, the traffic from the depicted users can be offloaded from the central storage. Associating all users to their strongest node will provide each of them with the best radio conditions. If there are no

issues with regards to congestion in the backhaul connection or the wireless channel, this is the optimal user assocation.

If, however, the backhaul connections are prone to congestion due to a lack of capacity, the cache at node 1.1 becomes less useful. Consider, thus, as a second example, caches installed at the BSs 2.1 to 2.4. Then, some users can be served from these edge caches and offload traffic from the backhaul connections. The MNO's user association policy now becomes important.

- All users are associated to the BS that provides the strongest signal. In the context of this work, such user association policy is called *closest BS* or simply CLOSEST. Then, to achieve 100% hit ratio, 6 cache units are required (assuming unit-sized content). If only 4 cache units are available, a hit ratio of no more than 2/3 can be accomplished.
- Cache-aware association. If users can be associated to any covering BS, a hit ratio of 100% can be achieved already with 4 cache units.

1.6 Scope of the Thesis and Contributions

In this thesis, three core questions of edge caching are addressed: *cache dimensioning and economics*, *content placement* and *user association*. The focus lies on the *economics* of edge caching in a wireless cellular network. Caches are installed at the BSs by the MNO. The BSs may or may not be equipped with high-speed backhaul links and might have different ranges, but are assumed to use equivalent radio access technology. The MNO offers the available edge cache commodity to CPs for financial compensation (*leasing* for fixed price or *price-adaptive*). CPs are assumed to take rational decisions with regards to the acquired cache space as well as *content placement*. *Entire files* are placed in a *prefetching* phase (in off-peak hours) since high-speed backhaul connections might not be present. For good placement decisions, *spatio-temporal traffic and popularity data* are assumed to

be given. The significance of MNO *user association* policies, cache-aware or not, is a center-piece to the CP decisions. The underlying assumption is that users in coverage overlaps may potentially be associated to any of the covering stations, taking the trade-off between local content availability and radio conditions into account. Mobility is only implicitly taken into account, e.g. by assuming a correlation of content popularity in neighboring areas.

In particular, the contributions of this thesis are:

- This thesis proposes and analyzes a **business model** for the **leasing** of edge caches from MNOs to CPs. This way, economic incentives of edge caching are considered that are a condition to the viability of the concept.
- **Cache-aware user association policies** are studied. When users in the overlap of coverage areas may be associated to any covering CBS, they potentially can access the content of more than one cache. The thesis introduces new cache-aware user association policies, trading off signal strength and availability of cached content. They can achieve different objectives such as load-balancing or throughput maximization. It is shown that these policies improve the perfomance of edge caching.
- Based on cache-aware user association, this thesis to the best of the author's knowledge is the first to **jointly optimize cache leasing**, **content placement** and **user association** decisions based on fine grained traffic and content popularity data and assuming that entire files are cached. The optimal **pricing** policy is determined, making the caching business most profitable for the MNO.
- The influence of **competition** between multiple CPs on cache pricing is explored.

1.7 Structure of the Thesis

The remainder of the thesis is structured as follows.

Chapter 2 provides an overview over the relevant literature on the wireless edge caching as it relates to the edge caching challenges outlined in Section 1.4.

Chapter 3 identifies the interests and incentives of the actors that participate in edge caching. The economic as well as technological conditions of edge caching are taken into account. A business model is proposed in which an MNO gives cache access to one or several CPs. A CP should rationally weigh the benefits from edge caching, e.g. traffic offloading from its own content storage, against the financial costs. It is examined how the MNO's user association policy is central to the CP decision.

Chapter 4 investigates user association in edge cache-equipped wireless networks. The problem of associating users to stations with given cached content is modelled as a (generally) non-linear Network Utility Maximization (NUM) problem by introducing a utility function per station. Choosing the utilities as concave functions may optimize the hit ratio while putting soft limitation on the served user load. Load-balancing among CBSs is of importance since the radio resources are limited, particularly when it is assumed that some CBSs can only serve a small number of users. The addition of weights for user-CBS associations allows for performance criteria such as system throughput.

A novel algorithm to optimally solve this problem in a distributed way is developped, called *(Generalized) Bucket-filling*. The calculations can be executed on the individual stations requiring a limited amount of information exchange. The resulting user association is shown to improve the desired

objective over cache-oblivious association policies.

Chapter 5 In Chapter 5, the work focuses on the problem of cache leasing and content placement taking user association into account. A mixed integer NUM problem is formulated that aims to maximize the CP's caching benefits given the user association policy.

As solution technique, *Generalized Benders decomposition* of the NUM into Master and Slave sub-problems is applied which converges to the global optimum. One of its main advantages, aside optimality, is that it allows the separation of the user association problem (Slave) from the cache leasing and content placement problem (Master) in an iterative solution process. The technique is completely original for solving NUM edge caching problems with non-linear utilities.

Extensive evaluation of the optimal leasing and content placement for linear and concave objective functions is provided. The results are compared to known content placement policies under cache-aware and cache-oblivious user association.

The evaluations can derive the optimal price that the MNO should set for its revenue maximization. Furthermore, they can suggest appropriate investment budget for the CP to attain a target hit ratio.

Chapter 6 develops a model for the allocation of cache space to several CPs. As preliminary results, the model is restricted to equal partition of cache space at each CBS with relaxed integrality constraint. A game is designed that helps allocate the memory even when the CPs anticipate MNO pricing. An algorithm towards a high quality Nash equilibrium is outlined.

Chapter 7 concludes the work. Several possible future extensions of the work are discussed.

The papers published during to this thesis can be found **the Appendix B**.

Chapter 2

Related Literature

2.1 Cache Economics

The business aspects of edge caching with its potentials and limitations are comprehensively discussed by Paschos et al. [PBL+16]. Cache leasing schemes have been approached from a competitive [PMGEA16, MMPC17] and a cooperative [PIP+16, DEAH17] perspective. In [PIP+16], Poularakis et al. propose offloading of backhaul traffic to local caches jointly optimizing caching incentive, content placement and routing policies. However, their Lagrangian based solution does not converge to the global optimum due to weak duality (see p. 143 in [BW05]) and other solution techniques are necessary to solve the problem optimally. In [ADR16] and [DCNT17], the authors investigate blind cache splitting between several CPs. The partitioning of the caches remains under the control of the internet service provider, while the CPs are allowed to establish secure connections between caches and users. Liu et al. [LLS+17] develop a cache leasing system for small base stations within the framework of contract theory without considering optimal content placement. In the context of Information Centric Networking (ICN), joint cache partitioning and resource allocation has been proposed recently

[CDL+17].

2.2 Content Placement

Question: Which content should be placed into edge caches? There are two notions of how caches should be updated: *Online Caching* and *Prefetching*.

2.2.1 Online Caching

In 2002, Che et al analyzed hierarchical web caching systems with Poisson arrival rate and a Zipf-like content popularity [CTW02]. They derive an approximation (called *Che approximation*) for the expected time that content remains within the cache before being replaced under the LRU replacement policy, and subsequently approximate the hit ratio. This implies that LRU can be approximated by assigning an appropriate Time-to-live (TTL) to each cached file that is independent of the other files in the cache [FNNT12]. Another well-known cache replacement policy is Least Frequently Used (LFU) [MWZ+17].

Online caching policies are evaluated with regard to knowledge of arriving request sequences. LRU optimizes the hit ratio under adversarial finite request sequences [ST85]. When the request sequence is stationary, LFU optimizes performance by giving advantage to content that is requested with higher frequency [DMT+16]. With TTL policies, the hit rate of different files can be adjusted individually. More recent research focuses on time-varying content popularity [ER15, LSLS18] modelling eviction policies as Markov chains.

Many works have applied variations of LRU to edge caching. Giovanidis and Avranas [GA16] exploit coverage overlaps in wireless networks under spatially varying content popularity. The resulting caching policy is called *multi-LRU*. The hit ratio is computed using the Che approximation. Other

works pursuing online caching policies include [LPG+16, TAG+15, NCM17, PIP+16, DJS+17, DCNT17, GH17].

While the structure of the incoming content requests should be known in order to select the optimal replacement policy, online caching does not require knowledge about the specific content to be cached. If, however, the caches are not equipped with high-speed connections to the core network but rather are deployed at wireless helper cells, content placement policies are required that preempt what content should be cached.

2.2.2 Prefetching

A large body of research falls into the category of prefetching (a selection includes [SGD+13, BBD14, PT13, PIT14, WCT+14, BG15, NMB+15, ADR16, PIST16]). Within the prefetching literature, some works have considered probabilistic cache placement [BG15, AGS17] while others study it in a deterministic way [SGD+13, PIP+16, PIST16]. User association is either performed to the closest station [BBD14] or more involved policies are applied that allow users to access content that is cached in all covering CBSs [SGD+13, BG15, PIP+16, PIST16].

More specifically, the original FemtoCaching problem (Shanmugam et al. [SGD+13]) assumes a bipartite connectivity graph of potential user associations. The aim is to place content such that user delay is minimized. Baştuğ et al. [BBD14] randomly place CBSs on the plane. They use metrics of outage probability and average delivery rate to analyse the performance of caching the most popular content. Poularakis, Iosifidis, and Tassiulas maximize in [PIT14] the hit ratio by means of integer optimization. In their work, an approximation scheme for hit ratio maximization is provided. Naveen et al. [NMB+15] provide an optimal placement and user association scheme with fractional content placement. Deghan et al. [DSJ+15] also develop an approximation algorithm for the content placement problem minimizing net-

work delay. In [BG15], Błaszczyszyn and Giovanidis develop a probabilistic content placement policy which maximizes the hit ratio by profiting from multi-coverage. Tuholukova, Neglia and Spyropoulos [TNS17] investigate optimal content placement for joint transmission schemes in small cell networks. Alfano et al. [AGL16] include power control issues in this general problem.

Jaffres-Runser and Jakllari [JJ18] performed field experiments that showed significant periods in which users are not connected to WiFi and only have cellular connection available. With the aim to avoid cellular downloads, they develop a data-driven algorithm that individually predicts what and how much content to cache on the wireless equipment during WiFi time.

2.3 Routing and User Association

Starting from the original FemtoCaching paper [SGD+13], multi-coverage is exploited: When a user can be associated to more than one cache, the available cached content is the union of the content of the covering caches. Since users conventionally associate to the wireless node with the strongest signal, novel association techniques are required to take full advantage of multicoverage opportunities.

In the existing literature, user association is handled in different ways: In [BGW10], the authors use caching to minimize bandwidth cost in a tree-like network. The routing decisions of users are, however, independent of each other. The authors of [SGD+13] associate users to any covering station that caches their requested content without balancing the traffic loads. In other works [BG15, BBD15], users are associated to the closest base station, not knowing if the requested content is stored in the cache or not. The authors of [PIT14] maximize the hit ratio by means of integer optimization. They introduce a bandwidth constraint limiting the amount of users that can be

connected to each cellular station. In [DSJ+15], user association is balanced between a cached and an uncached path. Association to the individual caches is modeled by shortest distance, again not allowing control over the use of the separate resources. The model in [NMB+15] includes both fractional content placement and routing variables and allows for the balancing of user traffic loads at the cache-equipped base stations.

The discussed studies consider user association assuming unicast transmissions. The advantages of multicast and broadcast transmission schemes are taken into account particularly in works on coded caching (see Section 2.6). Tuholukova, Neglia and Spyropoulos [TNS17] investigate optimal content placement for joint transmission schemes in small cell networks. Other works on CoMP include [LBZL17] and [CLQK17]. Liu et al. [LLS+17] develop a distributed content placement algorithm for different transmission schemes.

2.4 Traffic and Popularity Measurement

In 2005, Newman [New05] proposed and verified the Zipf distribution to well approximate the hit distribution of Internet content. Many works on wireless edge caching are based on this distribution as the assumption for content popularity, e.g. [SGD+13, PT13, BBD14, WCT+14]. However, it has been shown that spatio-temporal differences in user behaviour do affect the efficacy of caching-related decisions [MWZ+17, BNGP17].

The relationship between the popularity and location of online videos is studied in [BSW12]. The study finds that Youtube videos are essentially local: More than 50% of Youtube videos have more than 70% of their total views in a single country. Furthermore, it is found that video popularity expands as a wave: Video views immediately grow in the focus location and only then they expand across other regions. Li et al follow a more fine-

grained approach studying how content popularity depends on the specific location within a big city such as Shanghai [LXL+14]. They find, for example, that the concentration of video popularity becomes higher as the region is closer to downtown. The prediction of video popularity based on historical information given by earlier popularity measures is proposed in [PAG13]. In [XvdSLL15], social media data are used to predict content popularity. User location and mobility statistics combined with mobile application usage and the history of past cache queries are exploited for popularity estimates in [LSLT16]. A transfer learning-based approach is proposed to obtain an estimate of the popularity profile [BNGP17]. Using ideas of Gibbs sampling, Chattopadhyay, Blaszczyszyn and Keeler [CBK18] learn content popularity profiles in order to derive optimal content placement.

2.5 Mobility

Building on the original FemtoCaching paper, Wang, Song and Han [WSH15] consider the constantly changing topology of wireless networks in terms of which users are covered by access nodes. The mobility-aware cache update problem is solved suboptimally. In [PT17], Poularakis and Tassiulas model user mobility as random walks. A mobility-adapted placement policy offloads traffic from a macro base station to cache-equipped small base stations. The authors of [GXF+14] follow a similar approach, naming their content placement policy "MobiCacher". Chen et al [CHH+17] add the aspect of green content delivery to mobility-aware content placement.

The opportunities of D2D communication of devices equipped with cache memory are exploited by Lan et al [LWHS15]. Traffic is offloaded from the macro base station through proactive caching on mobile devices. Wang et al [WZSL17] follow a similar approach. They find that very slow and very fast moving users should cache the most popular files while mid-speed users

should cache less popular content.

Mahmood et al [MCC+16] apply mobile device caching to connected cars. Gorcitz et al [GJS+12] propose the use of vehicles equipped with large caches for massive data transmissions. Lee et al [LGP+16] investigate Fog Computing opportunities in vehicular networks as Vehicular Fog.

2.6 Fractional Caching

Some works on wireless caching consider fractional caching, in other words: fractions of files are stored instead of entire files. In some cases, partial caching is considered since it simplifies the content placement problem particularly if it is formulated as a mathematical optimization problem (e.g. [NMB+15]).

From an information-theoretic point of view, the fundamental limits of caching have been analyzed making use of the idea that files can be partitioned into subfiles [MAN14]. Its basic premise is that broadcast opportunities can be exploited even if different users simultaneously request different files. Users cache subfiles on their devices in a prefetching phase following the *centralized coded caching* scheme that depends on the user configuration. In the delivery phase the users simultaneously request content. For any configuration of the requested files, a bitwise XOR combination exists from which each user can reconstruct its request. Transmitting this bit combination over the common (perfect) channel significantly saves wireless resources in comparison to uncoded file transfer through the *global caching gain*.

Coded caching is regarded in several works on wireless caching, mostly as a theoretical bound that is compared with uncoded caching gains, see for example [SGD+13].

While centralized coded caching can obtain optimal caching gains in theory, there are practical drawbacks: The problem is largely intractable, even

though there exist 12-approximations. More importantly, the caching scheme assumes that files can be partitioned into sets of subfiles depending on the configuration of users present. Furthermore, a known user configuration in the prefetching phase is in conflict with the mobility inherent in wireless communication.

2.7 Privacy

Caches can be managed and trusted by a CP, as is the practice of Netflix OCAs [PBL+16, Izr16] (see also Section A.3.2). In this case, the cache is enabled to be the endpoint of secure E2E connections. Network management such as cache-aware user association can in this case only be performed if at least one of the two ends, user or CP, actively contributes.

If the caches are not managed and/or trusted by the CP, several attacks on communication privacy are identified in the literature. Acs et al [ACG+13] find as a major threat to user privacy the possibility of an adversary to measure round-trip time (RTT) to a local cache that operates with dynamic caching. Since the RTT to the cache is shorter than the one to the remote server, the adversary can then infer if content has been recently requested, assuming a replacement policy such as LRU. Several techniques that provide tradeoffs between privacy and latency are proposed to maintain lower latency through caching with privacy protection. Mohaisen et al [MZS+13] follow a similar approach, extending the model from a single cache-equipped router to a multi-hop path. Lauinger et al [LLR+12] propose replacement strategies that avoid caching of privacy-sensitive content.

Leguay et al [LPQS17] develop a protocol that allows caching of encrypted content through pseudo-identifiers in order to balance privacy with caching performance. This technique opens the possibility for an MNO to identify and deliver encrypted files without knowing their content.

Chapter 3

Leasing and Managing Edge Caches

3.1 A Business Model

3.1.1 Stakeholders

The three main parties that are stakeholders in wireless content delivery are MNOs, CPs and users.

A **user** subscribes to the services of an MNO and pays a subscription fee that arises per time period (e.g. per month) or per transmitted data volume. When the user is within the coverage area of one of the MNO's BSs and requests some CP's content, he/she expects content delivery via the wireless channel with satisfactory QoS, e.g. low latency and high throughput. At the same time, a certain level of data privacy needs to be provided that is related to E2E encryption of the connection to the respective CP.

CPs offer content to users. In exchange, users may have to pay a subscription fee and/or generate CP income by accepting the CP displaying advertisement and collecting data about user behaviour. Apart from the investment into production or licensing of the actual content, CPs also need

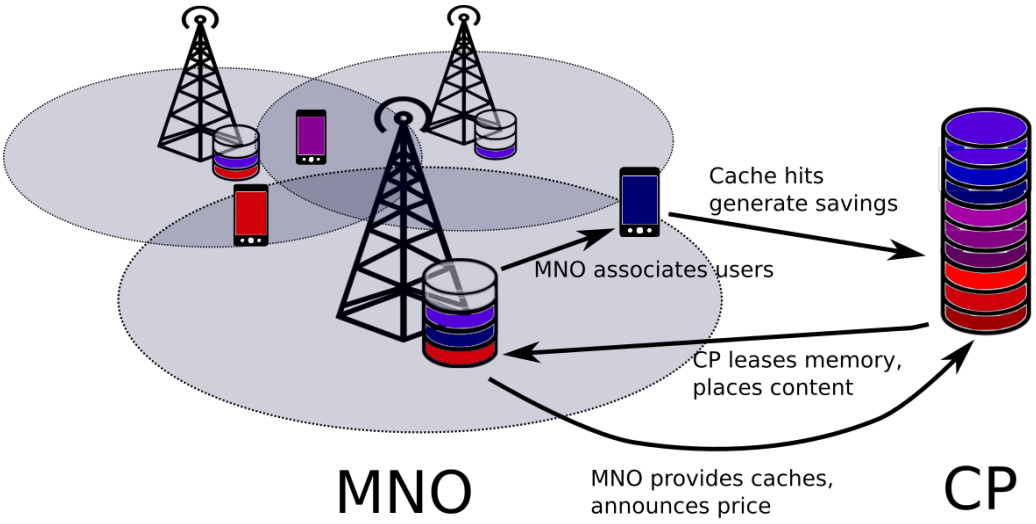

Figure 3.1: Cached wireless network run by the MNO, caches leased by the CP.

to pay for the storage of their content, be it in their own data centers or in CDNs. As has been shown above, an ever-increasing share of content delivery occurs to wireless users. This implies that CPs need to adapt to the wireless cellular networks to keep their customers satisfied.

An **MNO** builds and maintains a wireless cellular network. Its main source of income is user subscription fees. Since MNOs are in competition to each other, they have an incentive to provide good QoS to the users. This implies a good management of resources such as LTE resource blocks or backhaul capacity as well as sufficient investment into the network infrastructure to keep up with increasing demands.

3.1.2 Edge Cache Leasing

This thesis proposes the following business model for edge caching.

- An MNO installs cache memories at the BSs of its network, making them Cache-equipped BSs (CBSs). Computational capacity contolling each cache is installed locally at the caches. *The construction and maintenance of caches is left to the MNO since all hardware in a wireless*

network are maintained by the MNO.

- The memory space at the CBSs is to be **leased** to one or several CPs for a certain time period. The cache management is thus in the hands of a CP. This includes the capacity to establish a secure connection between the cache and a user. *The CPs are the owners of the content. Secure connections between user and CP can be provided when the CP can encrypt its content at the caches.*
- The MNO chooses a **pricing** mechanism, either a fixed price (in case memory is leased to one CP) or a price-adapting scheme (several CPs). *The CPs financially reward the MNO in return for the service of using the caches since edge caching a) generates savings for a CP by offloading traffic from its central content storage and b) provides better user QoS.*
- Once a CP has leased memory space, it can **place** its own **content** into them in a prefetching phase. The basis of the content placement decisions are spatio-temporal traffic and popularity data. *The CPs are the only entities that have traffic and popularity data to their disposal.*
- Content is transmitted to users in the delivery phase. Whenever a user is associated to a CBS that caches the requested content, it can be delivered directly from the cache. The MNO performs cache-aware or cache-oblivious **user association**. Cache-aware implies that, for users that may potentially be associated to more than one CBS, the placed content is taken into account. *Cache-aware user association brings edge caching to its full potential.*

This business model is represented in Figure 3.1.

3.1.3 MNO and CP decisions

Based on the previous discussion, we summarize the actions taken by the MNO and the CP. The MNO makes three strategic decisions:

MNO-1) How much storage space to install at each CBS?

MNO-2) In case of fixed price: Which price to set for the leasing of one cache unit? For multi CP scenarios: Which mechanism is to be used to allocate cache space to CPs?

MNO-3) Which user association policy to pursue?

The CP wants to lease sufficient cache space and place its content so that the savings exceed the cache leasing costs.

The two types of decisions that the CP takes are:

CP-1) How much cache space to lease from the MNO at each CBS?

CP-2) Which content to place into the leased caches?

For the decisions to be optimal, they need to be based on the estimated spatio-temporal popularity of the CP's content. Demand statistics evolve over time. The popularity data that are significant for the CP decisions are the aggregate data for the time window in which cache leasing and content placement decisions are fixed.

3.2 Interplay of User Association and Content Placement

The amount of savings generated by a cache leasing and content placement decision depends on the user association policy that the MNO declares. To understand this, we present the two perspectives:

a) Given a fixed content placement over all the caches in the CBSs, different user associations lead to different hit ratios. Consider a policy where the users are associated to the station with the strongest signal without taking content placement into account. Such association will result in a lower hit ratio than an alternative policy which considers cache placement information.

b) For a given user association policy, the CP can make leasing and placement decisions in a way that the hit ratio is maximized.

Instead of the hit ratio, other performance criteria can also be taken into account.

3.3 Performance Criteria

Cache-aware user association can be designed by various possible performance criteria.

The **hit ratio** or **hit probability** is the ratio of *cache-related* traffic, i.e. users that find their requested content cached, and total traffic. Since every cache-hit offloads traffic from the backhaul, the deeper network as well as the CP infrastructure, a maximum hit ratio is desirable. When user association aims to maximise hit ratio, users may suffer bad channel conditions.

This problem can be addressed by associate users to CBSs with the aim to **maximize** network **throughput**, i.e. the sum of the throughput achieved by each individual user. This criterion incentivizes to associate users to CBSs that provide good radio conditions together with the requested content.

Load balancing should be performed so that wireless nodes with popular content are not be overloaded with cache-related traffic. At the same time, other nodes with less popular content could be underused. A solution for such imbalance can be to limit the maximum cache-related load per wireless node and redistribute, when possible, the remaining load among neighboring nodes.

3.4 User Association

We have seen that user association plays a critical role in defining the approporiate CP decisions and in determining the system performance. This can be better explained through the example shown in Figure 3.2. In a random wireless network, unit-size content is placed into space-limited caches

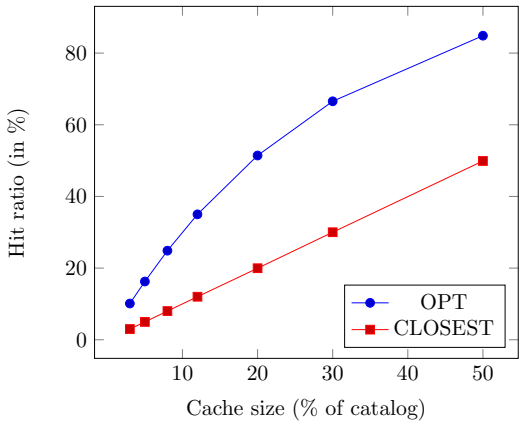

Figure 3.2: Hit ratio for RANDOM content placement policies with fixed cache sizes.

according the content placement policy RANDOM. This means that, at each cache, every k-subset of files uniformly has the same probability to be placed. The file popularity follows a Zipf distribution. Now we compare two different user association policies on the same content placement. OPT is cache-aware user association where users will be associated to any covering CBS caching their content, if such CBS exists. CLOSEST associates strictly to the closest station – or to the station that provides the best radio conditions. Fig. 3.2 now shows that OPT user association achieves a significantly higher hit ratio than CLOSEST. Particularly, this is due to the fact that users in coverage overlap areas gain access to all covering caches with OPT association. With CLOSEST, every user only has access to one cache.

3.5 User Association Policies

The following user association policies are considered in this work. They either do or do not take CP content placement decisions into account. Thus, they are either *cache-oblivious* or *cache-aware*.

Cache-oblivious Policy

Closest or *Standard Association*. The MNO associates each wireless user to the geographically closest CBS. This is the conventional policy in cellular networks that only takes signal quality (through distance) into account. In this case, users can be served locally only if their requested content is stored in that particular cache.

Cache-aware Policies

Closest Available or *Closest Replica*. Any user is associated to the closest covering CBS that caches the requested content.

Unsplittable or *Geographically determined association*. Users in the same geographic vicinity requesting the same content are associated to the same CBS. If users in an area are potentially associated to more than one station caching the content, it is selected apriori.

Opt- h. With Opt- h, association is a function of the placement decisions. Different performance criteria that determine Opt- h association decisions are discussed in Section 3.3. The h in Opt- h refers to a savings function that measures said performance criterion, as will be discussed later. Such MNO association policy allows for more user requests to be served by the caches. If users can be associated to any single wireless node among those with sufficient signal quality, then each user has potential access to the union of sets of files cached in all covering stations. The MNO should associate users to CBSs in such a way that content requests are matched to the cached content.

As an example for the different association policies, consider the following scenario that is also depicted in Figure 3.3:

Example 1. *A user is located in a wireless network such that he/she is covered by the CBSs L(eft), C(enter), and R(ight). His/her received signal is*

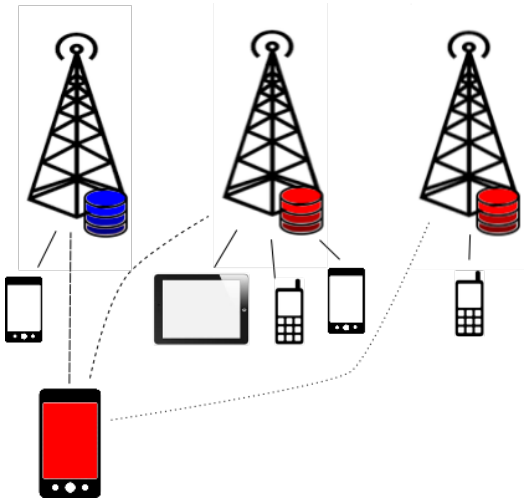

Figure 3.3: A user that can potentially be associated to three CBSs.

best from CBS L, but the requested content is not cached. The signal from C is weaker and several users are associated to it already. On the other hand, the requested content is cached. An even weaker (but still sufficiently strong) signal is received from R which also stores the content but has less traffic load.

With Closest *association, the user will receive the content from CBS L, even though the content has to be retrieved from the CP's upstream storage.* Closest Available *associates to C since it has the strongest signal of the CBSs caching the content. Possible congestion issues are not taken into account.* Opt-h *will associate to C or R since the content is stored in these caches. Depending on the performance criterion, the decision might be taken in favor of the better signal of C or the lower congestion at R.*

A further discussion on the practicalities of this matter follows in Section 3.7. The modelling and analysis of Opt-h vs Closest association in cached wireless networks is developped in Chapter 4.

3.6 Cache Leasing and Content Placement

Chapter 5 of this thesis develops a cache leasing and content placement policy that significantly improves on RANDOM content placement. It will be based on the knowledge of the MNO user association policy: When a CP knows how the MNO will associate users to CBSs, it can make its decisions in a way that maximizes its performance criteria. The performance criteria considered in this work are discussed in Section 3.3.

3.7 MNO-CP Cooperation

In order to make optimal leasing and content placement decisions, a CP relies on the MNO's pre-announced user association policy. Of particular interest is the policy OPT- h that associates users according to a performance criterion. In case OPT- h is applied, this thesis assumes that the MNO's and the CP's performance criteria for user association are the same. This should be part of the contract between MNO and CP.

OPT- h belongs to the cache-aware user association policies. Any cache-aware policy needs cooperation between the MNO and the CP. This is due to practical issues related to user privacy. Today's wireless systems do not allow an MNO to be aware of what content a user demands, and all information in encrypted as has been discussed in Section 1.4. However, the MNO is able to associate users in a cache-aware manner with guidance from the CP. There are several ways in which this can be implemented: The MNO can initially associate each user to a large set of covering stations but only serve the user from a CBS having the content. Alternatively, associate each user to a single CBS and redirect afterwards to a station with the content. In both cases, after initial association, the CP becomes aware of the user request and communicates the appropriate serving station to the MNO.

Chapter 4

Cache-aware User Association

This chapter contains research from the published papers [KGR17], [KGD18] and [KGDR18]. In Section 4.2, the model is based on [KGR17]. The weighted savings function (4.3) comes from [KGD18]. While the solution in 4.4 is based on [KGR17], the generalized solution presented here is developped in [KGDR18]. The evaluation for load balancing in Section 4.5 comes from [KGR17]. Section 4.6 is unpublished work.

4.1 Introduction

User association is a central topic in cache-equipped wireless networks as is shown in Section 3.4. This chapter analyzes the benefits of cache-aware user association. A CP has leased caches at the CBSs of an MNO's network and placed content into them. Users arrive in the network and request content. The MNO's task is now to associate the users to the CBSs. A basic assumption is that the bottleneck in wireless content delivery is the backhaul connection. For delay and backhaul congestion reasons, it is always preferable to associate users to a CBS that caches the requested content if it provides sufficent signal strength. Such traffic is called *cache-related*.

In practice, user requests arrive in real-time and the association decisions

need to be taken immediately. Here, aggregate user association is considered. The question is rather: "How many users will be associated to the CBSs after a time period, and to what effect?" rather than "Which CBS should the next incoming user request be associated to?". The main reason for this approach is that user association is seen here as a part of the cache leasing and content placement problem (CLCP, see Chapter 5). There, good cache leasing and content placement decisions are based on aggregate user traffic and popularity data. Thus, the *expected aggregate user association* needs to be calculated that follows the MNO's user association policy.

The large problem size when calculating user association for an entire cellular network makes centralized calculation of the optimal solution difficult. Following the idea that Fog computing capabilities are deployed with the caches, we develop a *distributed solution algorithm* for the user association problem that requires limited data exchange between the computing entities. This way, scalability of the solution is provided.

Different user association policies have been introduced in Section 3.5. This chapter examines the advantages of cache-aware user association Opt- h over the cache-oblivious policy Closest as well as other cache-aware varieties Closest Available and Unsplittable. Opt- h has its name since it optimizes user association with regards to a certain performance criterion (see Section 3.3), and each such criterion is measured by a related savings function h. The criteria that will be used in this chapter are hit ratio, load-balancing and throughput maximization.

4.2 System Model and Problem Statement

4.2.1 Network and Communications Model

Consider a cellular communications network with a finite set $\mathcal{M}$ of CBSs. A CP has leased cache space at the CBSs and placed content of the catalog

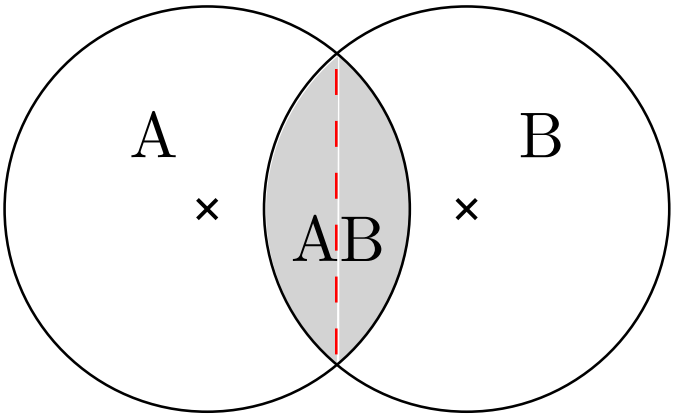

Figure 4.1: Toy example of network regions.

$\mathcal{F}$ into the caches apriori. The decisions for content placement were taken at the beginning of a time window and remain fixed throughout. They are expressed by the content placement vector $\mathbf{x} = (x_{m,f})$, $m \in \mathcal{M}, f \in \mathcal{F}$. If file f is stored at CBS m, then $x_{m,f} = 1$, otherwise $x_{m,f} = 0$. Note that the content placement $\mathbf{x}$ is a fixed parameter in this chapter. It is included in the optimization as a variable in Chapter 5.

Coverage Cells: The communications model is the following: Each CBS has a planar 2D coverage cell. Users covered by a CBS receive a radio signal strong enough to be potentially associated to it. Coverage cells may overlap, thus offering the users multiple options for service from covering CBSs. However, simultaneous service by more than one station, i.e. cooperative service, is not treated in this model, and left for future work.

Network Regions: The network area is partitioned into a set of regions. All positions in each region are assumed to experience the same radio conditions with respect to fading and interference. Furthermore, the MNO has a user association policy Π that allows for users in region s to potentially be associated to any CBS in $\mathcal{M}(s) \subseteq \mathcal{M}$, $|\mathcal{M}(s)| \geqslant 1$. ($\Pi$1) With the traditional Closest (see Section 3.5), $\mathcal{M}(s)$ is just the closest covering station. For the

OPT- h policy (see again 3.5), $\mathcal{M}(s)$ is the set of covering CBSs.

In general, for policy Π, the set of regions is denoted by $\mathcal{S}^{\Pi}$. A toy example with two CBSs is shown in Fig. 4.1: In case Π = OPT- h, there are three regions A, B and AB. Users in region A and B can only be associated to their uniquely covering CBSs, respecively. Users in region AB can potentially be associated to any of the two CBSs. If Π = CLOSEST (dashed line), there are two regions: A and the left part of AB contain traffic entirely associated to the left CBS, B and the right part of AB contain traffic belonging to the right CBS.

Content Popularity: For each region $s \in \mathcal{S}^{\Pi}$ and each content $f \in \mathcal{F}$, $N_{s,f}$ denotes the number of users in s requesting f. The content popularity vector is $\mathbf{N} = (N_{s,f})_{s\in\mathcal{S}^{\Pi},f\in\mathcal{F}}$. The statistics of content popularity are considered static during a time window. The information on file popularity is considered locally available at each covering CBS. Note that the model does not assume spatially uniform traffic and that the vector $\mathbf{N}$ is general.

In the context of this work, we are only interested in *the users who find their request cached at the CBS they are associated to (cache-hit traffic).* The association vector of *cache-hit* users to the CBSs is $\mathbf{y} = (y_{m,s,f})_{m\in\mathcal{M},s\in\mathcal{S}^{\Pi}(m),f\in\mathcal{F}}$, where $y_{m,s,f}$ represents the expected user traffic from region s requesting content f and associated with CBS m. $\mathcal{S}^{\Pi}(m)$ is the subset of regions whose users can potentially be associated to m according to Π. The vector $\mathbf{y}$ has fractional non-negative entries.

User association is unique in the sense that a single user cannot be served by two or more CBSs simultaneously. The total population $N_{s,f}$ can be distributed among the CBSs $\mathcal{M}(s)$, and some of it is potentially not associated to any CBS at all. Thus,

$$\sum_{m\in\mathcal{M}(s)} y_{m,s,f} \leqslant N_{s,f}, \ \ \forall\, s \in \mathcal{S}^{\Pi}, f \in \mathcal{F}. \tag{4.1}$$

This constraint allows for possible splitting of the population $N_{s,f}$ among the CBSs in $\mathcal{M}(s)$. The set of association vectors feasible to this constraint set is denoted by

$$\mathcal{Y}^{\Pi} := \left\{ \mathbf{y} \in \mathbb{R}_{\geqslant 0}^{\sum_{m\in\mathcal{M}} |\mathcal{S}^{\Pi}(m)||\mathcal{F}|} \;\middle|\; (4.1) \right\}.$$

Since we are only interested in cache-hit traffic, $y_{m,s,f}$ can only be nonzero if $x_{m,f} = 1$, i.e. if object f is cached in station m. Since no more than the total population requesting content f in s can be included in $y_{m,s,f}$, the following constraint set is valid:

$$y_{m,s,f} \leqslant N_{s,f} x_{m,f}, \;\; \forall\, m \in \mathcal{M}, s \in \mathcal{S}^{\Pi}(m), f \in \mathcal{F}. \tag{4.2}$$

This constraint set relates MNO association variables with CP cache placement decisions.

4.2.2 Savings Functions and Optimization Problem

As will be shown, every user association policy discussed here can be expressed as maximizing a savings function h under the aforementioned constraints. The savings function takes the user association vection $\mathbf{y}$ as input and maps it on the savings that can be achieved from it. Each of the savings functions is of the general form

$$\mathrm{h}(\mathbf{y}) = \sum_{m\in\mathcal{M}} \mathrm{U}_m(v_m^{\mathbf{w}}(\mathbf{y})), \tag{4.3}$$

where each U_m is a monotonously increasing, continuously differentiable and concave function that takes the weighted traffic volume at CBS m $v_m^{\mathbf{w}}(\mathbf{y})$ as its argument. Formally the latter is defined as

$$v_m^{\mathbf{w}}(\mathbf{y}) = \sum_{s\in\mathcal{S}^{\Pi}(m)} \sum_{f\in\mathcal{F}} w_{m,s,f} y_{m,s,f}, \tag{4.4}$$

with weights $w_{m,s,f}$ that are specific to users from region s requesting content f associated to CBS m. The vector of weights is $\mathbf{w}$. The unweighted cache-related traffic volume at m, in which all $w_{m,s,f} = 1$, is denoted by $v_m(\mathbf{y})$.

For a content placement vector $\mathbf{x}$ and a user association policy Π, the user association vector $\mathbf{y}^{\Pi}(\mathbf{x})$ is found by solving the following problem:

$$\text{(UA-}\Pi\text{)} \qquad \mathbf{y}^{\Pi}(\mathbf{x}) = \arg\max_{\mathbf{y}\in\mathcal{Y}^{\Pi}} \quad \mathrm{h}(\mathbf{y}) \qquad \text{s.t.} \quad (4.2).$$

Note that, due to the restrictions on h given in (4.3), UA-Π is a convex optimization problem in the context of network utility maximization (NUM).

4.3 Savings Functions

While this formulation is applicable to all user association policies that are taken into account, it particularly refers to the Opt- h policy where the savings function mirrors the different performance criteria (for introduction of association policies see Section 3.5). Before discussing the appropriate savings functions for the different performance criteria of user association policy Opt- h, we show that the optimization formulation UA-Π can also derive the user associations Closest and Unsplittable.

Policy i: Closest

For user association policy Π = Closest, finding the user association $\mathbf{y}^{\Pi}(\mathbf{x})$ is trivial. The solution is simply the covered users within the Voronoi cell of each CBS that request cached content. This can be expressed as the solution of the problem UA-Π by choosing the savings function $\mathrm{h}(\mathbf{y})$ proportional to the total sum of cache-related traffic:

$$\mathrm{h}(\mathbf{y}) = \sum_{m\in\mathcal{M}} v_m(\mathbf{y}).$$

For every region, all users located in it can only be associatiated to one unique CBS. Choosing the savings function this way assures that all users are associated that can find their requested content. Note that users that do not find their content cached are not counted into savings due to constraint (4.2).

Policy ii: Unsplittable

With the UNSPLITTABLE policy, there are regions in the overlap of several coverage areas. For each region and every content, association to a covering CBS is chosen apriori in case one of them caches the content. This association can be induced as the result of the problem UA-Π by introducing weights $w_{m,s,f}$ for each CBS m, the corresponding covered regions s and content f. Choosing

$$\mathrm{h}(\mathbf{y}) = \sum_{m \in \mathcal{M}} v_m^{\mathbf{w}}(\mathbf{y})$$

and $w_{m,s,f} > 0$ if users from region s requesting f are associated to CBS m while $w_{m,s,f} < 0$ leads to the association vector of the UNSPLITTABLE policy.

Policy iii: Hit Ratio Maximization

To maximize the hit ratio with OPT- h, the savings function can be chosen simply as the sum of the entries in the (unweighted) association vector $\mathbf{y}$:

$$\mathrm{h}(\mathbf{y}) = \sum_{m \in \mathcal{M}} \sum_{m \in \mathcal{M}} v_m(\mathbf{y}). \tag{4.5}$$

Policy iv: Load Balancing

When the objective of OPT- h user association, the aim is both to put a soft limit on the maximum traffic associated with a CBS and balancing the cache-related traffic volume among the CBSs. This approach also guarantees the *usefulness* of each cache.

We can choose the utility function such that for volumes greater than a certain V_m, the derivative of U_m becomes close to zero. Such soft limitation entails that it is not beneficial to further route users to this station. The overall objective is to maximize the sum of utilities subject to routing constraints, resulting in the savings function

$$\mathrm{h}(\mathbf{y}) = \sum_{m \in \mathcal{M}} \mathrm{U}_m\Big(\sum_{m \in \mathcal{M}} v_m(\mathbf{y})\Big). \tag{4.6}$$

Traffic association is balanced if the available resources are used in a fair way. Some notions of fairness are max-min, $\alpha-$ and proportional fairness. Each of them is achieved by appropriate choice of the utility functions (see [Kel97][MW00]). E.g. for proportional fairness, utilities could be chosen as (weighted) logarithms, depending also on the soft limit we want to achieve.

Policy v: Wireless Throughput

The weights can represent the downlink throughput between a user and a CBS (other work with the same objective is [LBZL17]). For such a weights choice, the channel quality between m and s is a constant value $h_{m,s}$ that depends on a reference distance and the path loss exponent. The emitted power level of m is denoted by p_m and the noise level by σ^2. Then, the signal-to-interference-plus-noise ratio (SINR) of users in region s when associated to covering CBS m is

$$\mathrm{SINR}_s(m) = p_m h_{m,s} \Bigg(\sum_{\substack{\tilde{m} \text{ covers } s \\ \tilde{m} \neq m}} p_{\tilde{m}} h_{\tilde{m},s} + \sigma^2 \Bigg)^{-1}, \tag{4.7}$$

where we assume that the interference from CBSs not covering s is negligible. For the downlink transmission from CBS m to region s, the throughput is equal to

$$w_{m,s} = B \log_2(1 + \mathrm{SINR}_s(m)) \quad \text{[in bits/sec]} \tag{4.8}$$

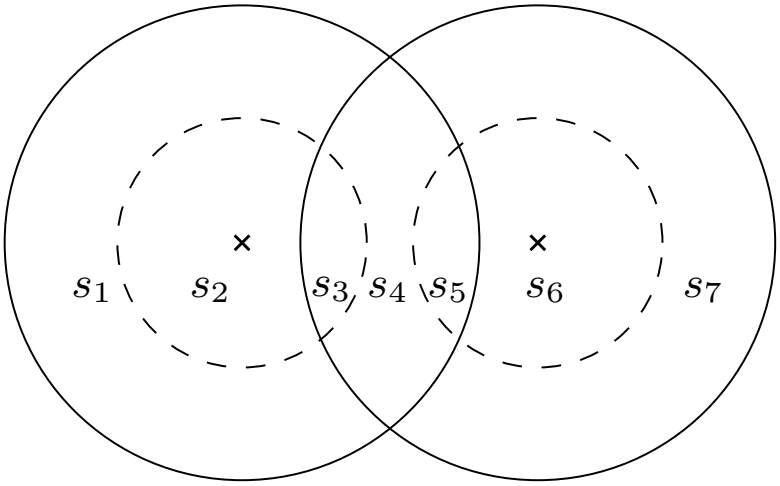

Figure 4.2: A network with refined regions consisting of two CBSs.

where B [Hz] is a chunk of bandwidth allocated to each served user. The total service bandwidth per CBS is equal to the product of B [Hz] times the users routed to the CBS. Using the value (4.8) as weights, UA-Π takes into account that it is favorable for a CBS to serve users with good radio conditions in order to use its resources effectively. Note that, we can also define other weights that depend on file f, for example $w_{m,s,f} = g(w_{m,s}/b_f)$, where $g(\cdot)$ is some increasing function and b_f is the file size. Such expression would then evaluate throughput over the requested file-size, giving larger weight w to (and thus favoring) smaller file sizes.

Fig. 4.2 shows an example of network regions with corresponding wireless performance weights based on the SINR. The circular areas with solid line depict the areas around each CBS where the received signal emitted by this station is above a certain threshold. For each CBS, we choose here to differentiate between two zones of signal strength (strong, weak), separated by the dashed lines. In the regions where there is disc overlap, users experience *interference* from the CBS they are not associated to. We can identify in this way 7 different regions with different overlaps. The SINR represented in the weights $w_{m,s,f}$ value the signal of m as beneficial and the other's as

interference.

It should be emphasized that the arbitrarily many levels of signal strength on the coverage area of each station can be introduced by appropriately re-defining the area partition $\mathcal{S}$. The modelling tradeoff is between the precision of communications aspects and the runtime of the optimization process.

The overall savings function is then given as

$$\mathrm{h}(\mathbf{y}) = \sum_{m \in \mathcal{M}} \sum_{m \in \mathcal{M}} v_m^{\mathbf{w}}(\mathbf{y}). \tag{4.9}$$

Other Performance Criteria

There are is a wide array of other performance criteria that can be applied to this model. Examples are:

a) Prioritized Caching: If the weights $w_{m,s,f}$ are proportional to the file sizes b_f, files of larger size that create more burden to the backbone are favorized to be cached. If they are inversely proportional, files of smaller size are preferred. Another work with similar objectives is [NCM17].

b) FemtoCaching: The users in [SGD+13] are equivalent to the regions as defined in our work. By choosing $w_{m,s,f}$ as the delay weights (see [SGD+13, maximization (3)]) together with zero leasing costs, the user association in FemtoCaching can be subsumed in out more general framework.

4.4 Solution

The user association problem UA-Π is a non-linear fractional optimization problem. As such, there are standard methods, such as gradient descent, that can solve it "out of the box". Here, we develop an algorithm to optimally solve the problem in a distributed way. The calculations can be executed on the individual stations requiring a limited amount of information exchange, thus making use of the Fog computing capabilities that come with the caches.

4.4.1 Simplification of the Problem

As a first step, we simplify the notation. Tuples of region s and file f are called *region-files* $q = (s, f)$ if there is a CBS m covering s with f in its cache, i.e. $\mathcal{Q}(m) \coloneqq \{(s, f) \mid s \in \mathcal{S}(m), f \in \mathcal{F}, x_{m,f} = 1\}$. The index q will be used as equivalent to the double index s, f in this section, e.g. we write N_q instead of $N_{s,f}$. We denote the complete set of region-files by $\mathcal{Q}$. Not included in $\mathcal{Q}$ are requests for files not stored in a covering CBS. The subset of CBSs that cache content f and cover region s is denoted by $\mathcal{M}(q) \subseteq \mathcal{M}$. Conversely, $\mathcal{Q}(m) \subseteq \mathcal{Q}$ are the region-files which can be served by the cache of m. Note that this definition renders constraint (4.2) redundant for $y_{m,s,f} = 0$ for all $(s, f) \notin \mathcal{Q}(m)$.

Recall that utility functions U_m are monotonously increasing. Therefore, in an optimal solution of the routing problem, all traffic of $q \in \mathcal{Q}$ is served from the cache of a CBS: For each region-file $q = (s, f) \in \mathcal{Q}$ there is a station covering s with f in its cache. As a consequence, all user requests related to region-files in $\mathcal{Q}$ (and only they) are cache-related traffic. Thus, constraint (4.1) is always fulfilled with equality in the optimum. Overall, UA-Π is equivalent to the problem

$$\begin{aligned} \max_{\mathbf{y} \geqslant 0} \quad & \mathrm{h}(\mathbf{y}) \\ \text{s.t.} \quad & \sum_{m \in \mathcal{M}(q)} y_{m,q} = N_q, \ \ \forall\, q \in \mathcal{Q} \end{aligned}$$

with $y_{m,s,f} = 0$ for all $(s, f) \notin \mathcal{Q}(m)$.

4.4.2 Dual method for the Augmented Lagrangian

The UA-Π problem is solved using the dual method on the Augmented Lagrangian (see [BT89], Section 3.4.4). We use the Augmented instead of the regular Lagrangian to achieve a distributed solution. In our case, the regular Lagrangian is not appropriate since it is not strictly concave in the primal

variables and hence the primal solution is not unique. This creates conflicts when different stations compete for the same users and convergence cannot be guaranteed. Like the regular Lagrangian, the Augmented one relaxes constraints of the UA-Π problem and introduces a *price* λ_q for the violation of each constraint. The difference between them is an additional quadratic term penalizing the violation of each constraint together with a factor $\varrho > 0$. This penalty guarantees strict concavity in the primal variables. Denoting the Augmented Lagrangian by $L^{(\varrho)}$, we get

$$\begin{aligned} L^{(\varrho)}(\mathbf{y}, \boldsymbol{\lambda}) = & \sum_{m \in \mathcal{M}} \mathrm{U}_m(v_m^{\mathbf{w}}(\mathbf{y}_m)) \\ & + \sum_{q \in \mathcal{Q}} \lambda_q (N_q - \sum_{m \in \mathcal{M}(q)} y_{m,q}) \\ & - \frac{\varrho}{2} \sum_{q \in \mathcal{Q}} (N_q - \sum_{m \in \mathcal{M}(q)} y_{m,q})^2, \end{aligned} \tag{4.10}$$

where $\boldsymbol{\lambda} \coloneqq (\lambda_q), q \in \mathcal{Q}$ is the price vector. The domains of the dual variables are $\lambda_q \in \mathbb{R}$ for all $q \in \mathcal{Q}$, since the respective constraints are equalities.

The Duality theorem (see [BT89], Appendix C) applies, which means that the duality gap is 0, and the dual method can be used. The objective function of the dual problem is

$$D^{(\varrho)}(\boldsymbol{\lambda}) \coloneqq \max_{0 \leqslant \mathbf{y}_m \leqslant \mathbf{N}_m, m \in \mathcal{M}} L^{(\varrho)}(\mathbf{y}, \boldsymbol{\lambda}) = L^{(\varrho)}(\mathbf{y}^*(\boldsymbol{\lambda}), \boldsymbol{\lambda}),$$

where

$$\mathbf{y}^*(\boldsymbol{\lambda}) = \operatorname*{arg\,max}_{0 \leqslant \mathbf{y}_m \leqslant \mathbf{N}_m, m \in \mathcal{M}} L^{(\varrho)}(\mathbf{y}, \boldsymbol{\lambda}) \tag{4.11}$$

is the primal maximum of (4.10) for a given price vector $\boldsymbol{\lambda}$. The dual problem is then defined as

$$\text{(UA-dual)} \qquad \min_{\boldsymbol{\lambda} \in \mathbb{R}^{\mathcal{Q}}} D_\varrho(\boldsymbol{\lambda}).$$

Starting from an arbitrary initial dual vector $\boldsymbol{\lambda}(0)$, the dual vector is iteratively updated according to

$$\lambda_q(t+1) = \lambda_q(t) - \varrho\Big(\sum_{m\in\mathcal{M}(q)} N_q - y^*_{m,q}(\boldsymbol{\lambda}(t))\Big), \tag{4.12}$$

where the steplength $\varrho > 0$ is the penalty used in (4.10). The convergence of this method is well known (see [Rus95] or Section 3.4.4 of [BT89]).

For practical implemetation issues of each update step of the region-file price $\lambda_q, q \in \mathcal{Q}$, only the primal solutions of the covering CBSs need to be known. Thus, for a distributed implementation, exchange of such information among neighboring stations is sufficient.

The next subsection presents the distributed solution for the primal problem (4.11), which needs to be found for every iteration of the dual algorithm.

4.4.3 Distributed solution for the primal problem

The solution for (4.11) is unique since the domain of $\mathbf{y}$ is convex and compact and, for any fixed feasible vector $\boldsymbol{\lambda}$, the Augmented Lagrangian $L^{(\varrho)}$ is strictly concave. We use the Diagonal Quadratic Approximation Method (DQA) [Rus95] to derive seperate problems which can be solved by each cache. A limited amount of exchanged information between neighboring caches is required.

The DQA overcomes the problem that the objective function $L^{(\varrho)}(\mathbf{y}, \boldsymbol{\lambda})$ of (4.11) is not easily separable among the variables related to the different CBSs, since it contains quadratic terms combining different variables $y_{m,q}$ (see (4.10)). To achieve this, we introduce the functions $L^{(\varrho)}_m : \mathbb{R}^{Q_m} \times \mathbb{R}^{\sum_{\tilde{m}} Q_{\tilde{m}}} \times \mathbb{R}^{Q} \to \mathbb{R}$ for all $m \in \mathcal{M}$:

$$\begin{aligned} L^{(\varrho)}_m(\mathbf{y}_m, \tilde{\mathbf{y}}, \boldsymbol{\lambda}) \coloneqq \mathrm{U}_m(v^{\mathbf{w}}_m(\mathbf{y}_m)) &+ \sum_{q\in\mathcal{Q}(m)} \lambda_q y_{m,q} \\ &- \frac{\varrho}{2} \sum_{q\in\mathcal{Q}(m)} (\bar{N}^m_q(\tilde{\mathbf{y}}) - y_{m,q})^2, \end{aligned}$$

where $\bar{N}_q^m(\tilde{\mathbf{y}}) := N_q - \sum_{\substack{\bar{m}\in\mathcal{M}(q)\\ \bar{m}\neq m}} \tilde{y}_{\bar{m},q}$ is the number of requests in q not associated with caches other than m in the routing vector $\tilde{\mathbf{y}} = (\tilde{y}_{m,q}), m \in \mathcal{M}, q \in \mathcal{Q}(m)$ which is here seen as a parameter. The primal problem to be solved by each cache m is defined as

$$\text{(UA-primal-}m\text{)} \quad \max_{0\leqslant \mathbf{y}_m \leqslant \mathbf{N}_m} L_m^{(\varrho)}(\mathbf{y}_m, \tilde{\mathbf{y}}, \boldsymbol{\lambda}). \tag{4.13}$$

Since $L_m^{(\varrho)}(\mathbf{y}_m, \tilde{\mathbf{y}}, \boldsymbol{\lambda})$ is strictly concave in $\mathbf{y}$ and the domain is compact, UA-primal-m has a unique solution which we call $\tilde{\mathbf{y}}_m^*$. The vector containing the solutions of UA-primal-m for all caches is $\tilde{\mathbf{y}}^*$.

The DQA method consists of parallel execution of UA-primal-m at the caches with consecutive update of the vector $\tilde{\mathbf{y}}$ in the fashion of a nonlinear Jacobi algorithm. It produces a succession of vectors $\tilde{\mathbf{y}}(0), \tilde{\mathbf{y}}(1), \tilde{\mathbf{y}}(2), \ldots$. Starting from some given vector $\tilde{\mathbf{y}}(0)$, the vector $\tilde{\mathbf{y}}(\tau+1)$ is defined as the convex combination of $\tilde{\mathbf{y}}(\tau)$ and $\tilde{\mathbf{y}}^*(\tau)$. Given a constant $0 < \alpha \leqslant 1$, we get

$$\tilde{\mathbf{y}}(\tau+1) = \tilde{\mathbf{y}}(\tau) + \alpha(\tilde{\mathbf{y}}^*(\tau) - \tilde{\mathbf{y}}(\tau)). \tag{4.14}$$

In [Rus95] it is shown that the DQA method converges. Observe that the convergence depends on the uniqueness of the primal solutions $\tilde{\mathbf{y}}^*(\tau)$. For every update (4.14), each station only requires results from its neighboring stations that cover a common region-file.

4.4.4 Separated Primal Solution

Before stating the separated primal problem, we again simplify the notation for this section. Since the problem is separated by CBS, we omit the index m, writing y_q instead of $y_{m,s,f}$, $\mathbf{y}$ instead of $\mathbf{y}_m$, and $\mathcal{Q}$ instead of $\mathcal{Q}(m)$. The separated primal problem is then

$$\begin{aligned}\mathbf{y}^* &= \operatorname*{arg\,max}_{0\leqslant\mathbf{y}\leqslant\mathbf{N}} \mathrm{U}\Big(\sum_{q\in\mathcal{Q}} w_q y_q\Big) + \sum_{q\in\mathcal{Q}} \lambda_q y_q - \frac{\varrho}{2}\sum_{q\in\mathcal{Q}} (\bar{N}_q - y_q)^2 \\ &= \operatorname*{arg\,max}_{0\leqslant\mathbf{y}\leqslant\mathbf{N}} \mathrm{U}\Big(\sum_{q\in\mathcal{Q}} w_q y_q\Big) - \sum_{q\in\mathcal{Q}} \big[(\varrho/2) y_q^2 - a_q y_q\big] \end{aligned} \tag{4.15}$$

where $\bar{N}_q = (\bar{N}_q^m(\tilde{\mathbf{y}})$ is the remaining user population not served by other CBSs in the previous DQA step, and $a_q \coloneqq \lambda_q + \varrho\bar{N}_q$. The equation to (4.15) comes from the development of the quadratic term and from omitting the additive constants which do not affect the optimal choice of values for the variables. In the following, we will denote the objective function of (4.15) by $g(\mathbf{y})$.

Problem (4.15) is a convex optimization problem. Methods for solving this type of problem such as the gradient descent method are well known. However, their speed of convergence can be an issue. Here, we present a problem specific solution method that does not depend on convergence and solves (4.15) exactly and efficiently. Our method is based on the insight that the optimal solution lies within a one-dimensional subspace of its $|\mathcal{Q}|$-dimensional domain. The following theorem characterizes this subspace.

Proposition 1. *Let* $\mathbf{y}^*$ *be defined as in* (4.15)*. Then, there exists* $\nu^* \geqslant 0$ *such that for all* $q \in \mathcal{Q}$

$$y_q^* = \begin{cases} 0, & \nu^* \leqslant \beta_q \\ N_q, & \nu^* \geqslant (w_{\tilde{q}}/w_q)N_q + \beta_q \\ f_q(\nu^*), & \textit{otherwise} \end{cases} \tag{4.16}$$

with

$$\beta_q = \frac{a_{\tilde{q}} - (w_{\tilde{q}}/w_q)a_q}{\varrho}. \tag{4.17}$$

In the above, $\tilde{q} \coloneqq \arg\max_{q\in\mathcal{Q}} a_q/w_q$ *and*
$f_q(\nu) = (w_q/w_{\tilde{q}})\,\nu - (w_q/w_{\tilde{q}})\beta_q$.

In other words, the optimal vector $\mathbf{y}^*$ can be determined by finding the optimal value ν^*. This fact considerably reduces the complexity of finding $\mathbf{y}^*$.

Before the formal proof of Proposition 1, consider the following illustration. The value $\nu \geqslant 0$ can be interpreted as the *water level* in a scenario

in which $|\mathcal{Q}|$ *buckets of varying width* $w_q/w_{\tilde{q}}$ and height $(w_{\tilde{q}}/w_q)N_q$ are positioned at different bottom levels. When the water level is below or exactly at bottom level β_q of bucket q, the bucket is empty. Otherwise, the bucket is filled up to level ν unless the water level is higher than the upper edge of the bucket at $(w_{\tilde{q}}/w_q)N_q + \beta_q$. In that case, the bucket is filled to its capacity of N_q. In between, the volume in the bucket is exactly $f_q(\nu)$. This observation leads to an algorithm solving (4.15) efficiently: Starting from water level $\nu = 0$, let the water level increase. The water volumes represent the variables in $\mathbf{y}$. The algorithm (see Algorithm 1 for detailed presentation) terminates when the optimal value ν^* is reached. This is true when the maximum of the objective function of (4.15) is reached or when all buckets are full.

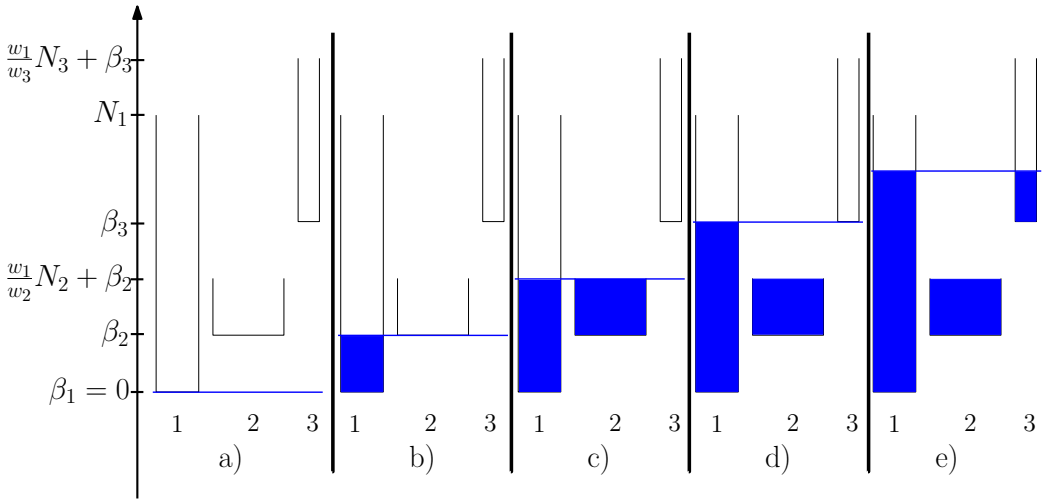

Figure 4.3: Illustration of the Bucket-filling algorithm.

An example for such an arrangement of buckets is shown in Fig. 4.3. Each bucket q is placed with its bottom at level β_q and has a width of $w_q/w_{\tilde{q}}$, a height of $(w_{\tilde{q}}/w_q)N_q$ and unit depth. The buckets are filled to a common level – or until they are full. The algorithm halts when further increasing the water level ν starts decreasing the objective value.

Proof of Proposition 1. Note first that $f_q(y)$ is invertible with

$$f_q^{-1}(y) = (w_{\tilde{q}}/w_q)y + \beta_q \tag{4.18}$$

and f_q^{-1} is non-decreasing. The value $f_q^{-1}(y)$ represents the reference water level ν if y is the non-zero water volume in bucket q. Let $\hat{y}_q := w_q y_q$ for all $q \in \mathcal{Q}$. Then, the following problem is equivalent to (4.15):

$$\hat{\mathbf{y}}^* = \underset{0 \leqslant \hat{\mathbf{y}} \leqslant \hat{\mathbf{N}}}{\arg\max} \quad \hat{g}(\hat{\mathbf{y}})$$

with $\hat{g}(\hat{\mathbf{y}}) := \mathrm{U}\left(\sum_{q\in\mathcal{Q}} \hat{y}_q\right) - \sum_{q\in\mathcal{Q}} \left[\frac{\varrho}{2w_q^2}\hat{y}_q^2 - \frac{a_q}{w_q}\hat{y}_q\right]$ and $\hat{\mathbf{N}} := (w_q N_q), q \in \mathcal{Q}$. Note that $N_q > 0$ for all q and thus the upper and the lower bound of y_q cannot be fulfilled with equality simultaneously. The KKT conditions can thus be written in a compact way: There exist $\gamma_q \in \mathbb{R}$ for all $q \in \mathcal{Q}$ such that

$$\frac{\partial}{\partial \hat{y}_q}\hat{g}(\hat{\mathbf{y}}^*) - \gamma_q = 0 \tag{4.19}$$

for all $q \in \mathcal{Q}$ where

$$\gamma_q = \begin{cases} \geqslant 0, & \text{if } \hat{y}_q^* = w_q N_q \iff y_q^* = N_q \\ \leqslant 0, & \text{if } \hat{y}_q^* = 0 \iff y_q^* = 0 \\ = 0, & \text{otherwise} \end{cases} \tag{4.20}$$

and

$$\frac{\partial}{\partial \hat{y}_q}\hat{g}(\hat{\mathbf{y}}) = \mathrm{U}'\left(\sum_{q\in\mathcal{Q}} \hat{y}_q\right) - \frac{\varrho}{w_q^2}\hat{y}_q + \frac{a_q}{w_q}.$$

For any $p, q \in \mathcal{Q}$, (4.19) implies that the following equation holds:

$$-\frac{\varrho}{w_q^2}\hat{y}_q^* + \frac{a_q}{w_q} - \gamma_q = -\frac{\varrho}{w_p^2}\hat{y}_p^* + \frac{a_p}{w_p} - \gamma_p$$

since the term $\mathrm{U}'\left(\sum_{q\in\mathcal{Q}} \hat{y}_q^*\right)$ is contained in both derivatives. A calculation resubstituting $y_p^* = \hat{y}_p^*/w_p$ and $y_q^* = \hat{y}_q^*/w_q$ results in $y_q^* = f_q \circ f_p^{-1}(y_p^*) +$

$(w_q/\varrho)(\gamma_p - \gamma_q)$. Now, we will find a $p \in \mathcal{Q}$ such that $\nu^* \coloneqq f_p^{-1}(y_p^*)$ yields (4.16). Then,

$$y_q^* = f_q(\nu^*) + \frac{w_q}{\varrho}(\gamma_p - \gamma_q) \tag{4.21}$$

Case 1: $\exists\, p \in \mathcal{Q}$ with $0 < y_p^* < N_p$. Define $\nu^* \coloneqq f_p^{-1}(y_p^*)$. Note that, in this case, $\gamma_p = 0$ such that for all $q \in \mathcal{Q}$:

$$y_q^* = f_q(\nu^*) - \frac{w_q}{\varrho}\gamma_q. \tag{4.22}$$

With this equation, we deduce (4.16) for any $q \in \mathcal{Q}$: Firstly, if $0 < y_q^* \leqslant N_q + \beta_q$, then $\gamma_q = 0$ follows from (4.20) and $y_q^* = f_q(\nu)$ holds due to (4.22). Secondly, if $y_q^* = 0$, then $\gamma_q \leqslant 0$ by (4.20). Then (4.22) implies that $f_q(\nu^*) \leqslant y_q^* = 0$ which by definition of f_q is true if and only if $\nu^* \leqslant \beta_q$. Lastly, if $y_q^* = N_q$, then $\gamma_q \geqslant 0$ by (4.20) and with (4.22): $f_q(\nu^*) \geqslant y_q^* = N_q$. By definition, this holds if and only if $\nu^* \geqslant (w_{\tilde{q}}/w_q)N_q + \beta_q$.

Case 2: $\nexists\, q \in \mathcal{Q}$ with $0 < y_q^* < N_q$, but $\exists q \in \mathcal{Q}$ with $y_q^* = 0$. Choose p such that $f_p^{-1}(0) = \beta_p =: \nu^*$ is minimal among all such lower-bound region-files, i.e. $p = \arg\min_{q \in \mathcal{Q}, y_q^* = 0} \beta_q$. It suffices to show that $f_q(\nu^*) \leqslant 0$ if $y_q^* = 0$ and $f_q(\nu^*) \geqslant N_q$ if $y_q^* = N_q$. Let firstly $q \in \mathcal{Q}$ with $y_q^* = 0$. Then

$$f_q(\nu) = f_q(\beta_p) = (w_q/w_{\tilde{q}})(\beta_p - \beta_q) \leqslant 0$$

by definition of p. Secondly, let $q \in \mathcal{Q}$ with $y_q^* = N_q$. Note that (4.20) implies that $\gamma_p \leqslant 0$ and $\gamma_q \geqslant 0$. Thus, (4.21) implies that $f_q(\nu^*) \geqslant N_q$ which holds if and only if $\nu^* \geqslant (w_{\tilde{q}}/w_q)N_q + \beta_q$.

Case 3: $y_q^* = N_q$ for all $q \in \mathcal{Q}$. Choose $p \coloneqq \arg\max_{q \in \mathcal{Q}} f_q^{-1}(N_q)$ and $\nu \coloneqq f_p^{-1}(N_p) =$. Then, since f_q is non-decreasing for $q \in \mathcal{Q}$,

$$f_q(\nu^*) \geqslant f_q(y_q^*) = f_q(N_q) = (w_{\tilde{q}}/w_q)N_q + \beta_q.$$

It remains to be shown that $\nu^* \geqslant 0$ in all cases. Note that $\nu = f_p^{-1}(y_p^*)$

Algorithm 1 Generalized Bucket-filling (solves UA-primal-m with objective function (4.3))

1: Choose $\tilde{q} \in \mathcal{Q}$ such that $a_{\tilde{q}}/w_{\tilde{q}} \geqslant a_q/w_q$ for all q.
2: Sort $\mathcal{Q}$ by β_q non-decreasingly.
3: Initialize the sets of *active* region-files as $\mathcal{Q}^A = \{q \in \mathcal{Q} \mid \beta_q = \beta_{\tilde{q}}\}$, *inactive* region-files $\mathcal{Q}^I := \mathcal{Q} \backslash \mathcal{Q}^A$ and *deactivated* region-files as $\mathcal{Q}^D := \varnothing$.
4: Set $\nu := 0$.
5: Increase ν until
- $\nu = \beta_q$ for some inactive $q \in \mathcal{Q}$, then $\mathcal{Q}^A := \mathcal{Q}^A \cup \{q\}$ and $\mathcal{Q}^I := \mathcal{Q}^I \backslash \{q\}$
- $\nu = (w_{\tilde{q}}/w_q)N_q + \beta_q$ for some active q, then $\mathcal{Q}^A := \mathcal{Q}^A \backslash \{q\}$ and $\mathcal{Q}^D := \mathcal{Q}^D \cup \{q\}$ or
- $\partial/\partial y_q \; g(\mathbf{y}) = 0$ for all $q \in \mathcal{Q}^A$, $\partial/\partial y_q \; g(\mathbf{y}) \leqslant 0$ for all $q \in \mathcal{Q}^I$, and $\partial/\partial y_q \; g(\mathbf{y}) \geqslant 0$ for all $q \in \mathcal{Q}^D$, where $\mathbf{y} = (y_q)$ with $y_q = f_q(\nu)$ for $q \in \mathcal{Q}^A$, $y_q = 0$ for $q \in \mathcal{Q}^D$ and $y_q = N_q$ for $q \in \mathcal{Q}^I$.

6: If the last condition is fulfilled, return $\mathbf{y}$. Otherwise, go to step 5.

for some $p \in \mathcal{Q}$, f_p^{-1} is non-decreasing (see (4.18)) and $y_p^* \geqslant 0$. Then

$$\begin{aligned} \nu = f_p^{-1}(y_p^*) &\geqslant f_p^{-1}(0) = (w_{\tilde{q}}/w_p)\beta_p \\ &= w_{\tilde{q}} \frac{a_{\tilde{q}}/w_{\tilde{q}} - a_p/w_p}{\varrho} \geqslant 0. \end{aligned}$$

The last step is true due to the choice of $\tilde{q} = \arg\max_{q \in \mathcal{Q}} a_q/w_q$. □

We can now describe the novel Generalized Bucket-filling algorithm that efficiently finds the optimal solution to UA-primal-m (see Algorithm 1).

Observe that step 2 can be done in $O(|\mathcal{Q}| \log(|\mathcal{Q}|))$ operations with a sorting algorithm such as quicksort. Thus, it dominates steps 1 and 3 that each need $O(|\mathcal{Q}|)$ operations. Step 5 is executed up to $2\,|\mathcal{Q}|$ times, since every region-file tuple is activated and deactivated no more than one time each. Assuming that f_q and $\partial/\partial y_q \; g(\mathbf{y})$ can be evaluated in $O(1)$, the runtime

of the loop in steps 5 and 6 is $O(|\mathcal{Q}|)$. This shows that the overall runtime is dominated by sorting in step 2 and thus is $O(|\mathcal{Q}| \log(|\mathcal{Q}|)) = O(\hat{S}\hat{F} \log(\hat{S}\hat{F}))$ where $\hat{S}$ is the largest number of regions covered by any CBS and $\hat{F}$ the largest number of files that can be cached by any CBS.

4.4.5 Algorithm

The complete algorithm that finds the globally optimal solution to the UA-Π problem is summed up in Algorithm 2.

Algorithm 2 Solve UA-Π

1: Choose dual vector $\boldsymbol{\lambda}(0)$, $t = 0$, $\varepsilon > 0$.
2: **while** $\|\boldsymbol{\lambda}(t) - \boldsymbol{\lambda}(t-1)\| > \varepsilon$ **do**
3: Choose $\tilde{\mathbf{y}}(0)$, $\tau = 0$.
4: **while** $\|\tilde{\mathbf{y}}(\tau) - \tilde{\mathbf{y}}(\tau-1)\| > \varepsilon$ **do**
5: Find $\tilde{\mathbf{y}}_m^*(\tau)$ with Algorithm 1 at every $m \in \mathcal{M}$ separately.
6: Exchange results among neighboring stations, set $\tilde{\mathbf{y}}(\tau+1)$ as in (4.14), $\tau = \tau + 1$.
7: **end while**
8: Exchange results among neighboring stations, set $\boldsymbol{\lambda}(t+1)$ as in (4.12) , $t = t + 1$.
9: **end while**

The choice of the first dual vector $\boldsymbol{\lambda}(0)$ in line 1 is arbitrary. The first primal vector $\tilde{\mathbf{y}}(0)$ in line 3 can be chosen as the last primal vector of the iteration before.

4.5 Numerical Evaluation for Load Balancing

At first, the model is evaluated with concave utility functions for load balancing purposes. Consider an urban area of 2.5 km × 2.5 km with uniform

user distribution. A catalog of 6 files of equal size is known. The popularity of the files follows a Zipf distribution with parameter 1. Throughout the simulations, CBSs are placed in the area following a Poisson Point Process (PPP) with density $8\frac{\text{CBS}}{km^2}$. This means that their total number in each run is a random Poisson realization, and their position is uniform in the simulation window. In this way we avoid the bias of testing only particular network topologies. We run two simulation scenarios, the first for single-tier and the second for two-tier netowrks. Each scenario consists of 1000 simulation runs and we consider the averaged results over the runs. Coverage follows the Boolean model where a disc area is centered on each wireless station with some defined radius. The surface of different overlapping areas is found in each run by the Monte-Carlo method. The users are routed to the CBSs following three different cache-aware policies:

1. The Opt- h policy from Algorithm 2 with logarithmic utility function for each cache (Policy iv from Section 4.3. This policy guarantees a proportionally fair (and also max-min fair) solution. We call this policy Fair.
2. The Closest Available policy, which associates each user with the closest CBS that both covers its position and has the requested content cached.
3. The Unsplittable policy which associates all users in a region requesting the same file with a unique random CBS among all covering CBSs having this content.

We want to evaluate the proportion of user traffic served by different CBSs over the total traffic routed to CBSs for each policy. In this way we can compare the policies based on how (un)equally they associate traffic load among the available CBSs. Observe that for all three policies, the total traffic volume associated to CBSs is the same, because in all scenarios the CBSs store the same cached content, and traffic is routed to a CBS whenever

possible. Hence, the comparison is fair.

4.5.1 Single-tier Networks

In an ideal situation, all stations would serve exactly the same amount of traffic. This, however, is normally not possible in a random network. A routing policy is better than another, when the maximum load of a CBS is lower and at the same time the minimum load share is higher than in the other policy. This way, an overload of the stations is avoided while the usefulness of all stations is achieved. Simulating single-tier networks, we want to verify that the fair policy provides a more balanced distribution of traffic to CBSs than the other two. The coverage radius of the CBSs is varied between 62.5 m and 500 m. This can be translated to an expected number of covering CBSs per user between 1 and 6. This mapping comes from the Boolean model [BG15]. Two different sets of content files with different popularities are placed uniformly randomly into the caches. Since we are only interested in traffic associated with CBSs, we disregard users not covered by any CBS.

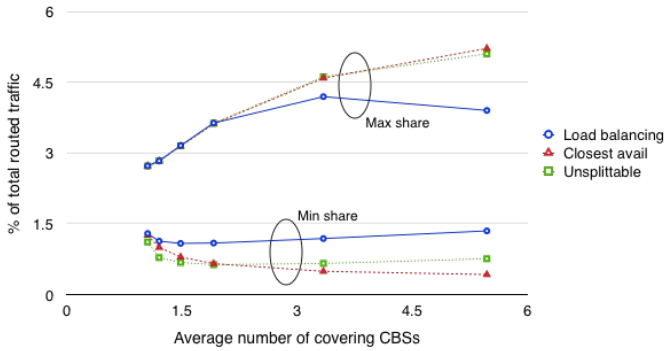

Figure 4.4: Minimum and maximum load share of a CBS in the network depending on the mean coverage number.

Fig. 4.4 illustrates how the routing decisions of each of the three policies affect the distribution of load shares among all CBSs. It displays the average maximum (upper curve) and minimum load share (lower curve) over the

mean number of CBSs a covered user can see.

The results show that with an increasing average number of covering stations, the fair policy achieves a lower maximum load as well as a higher minumum load. Since the overall traffic routed to the CBSs is the same for all policies, we can conclude that the FAIR policy makes the most balanced use of the available resouces. The resources which remain available for potential cache-unrelated traffic are spread evenly across the network.

4.5.2 Two-tier Network

In a second scenario, we simulate an area covered by two tiers [DRGC13]: one of large and one of small coverage. We show that the fair policy is better at offloading traffic from larger to smaller CBSs than the other policies (see Fig. 4.5). While the FAIR policy burdens the small stations with a higher load, we demonstrate that it distributes the load more evenly among them so that no individual station is overburdened (see Fig. 4.6). The first tier consists of large CBSs having a 187.5 m coverage radius while the second tier of small CBSs has 62.5 m coverage radius. The large CBSs are equipped with caches and the two most popular files are stored in all of them. The smaller CBSs get one of these two files assigned uniformly randomly in their cache.

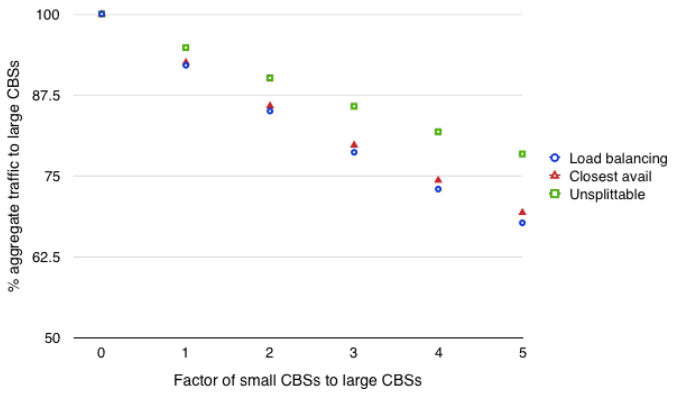

Figure 4.5: Aggregate traffic share of large CBSs depending on the amount of small CBSs.

In Fig. 4.5, we show the percentage of all traffic routed to large CBSs for each of the three policies. The x-coordinate increases with the ratio of small stations over large stations in the network. When less traffic is routed to the large CBSs, then the policy provides a more efficient offloading of traffic towards the small CBSs. The figure shows that, increasing the amount of small CBSs, the FAIR policy offloads significantly more traffic to the small stations than the UNSPLITTABLE policy, and slightly more than the CLOSEST AVAILABLE policy.

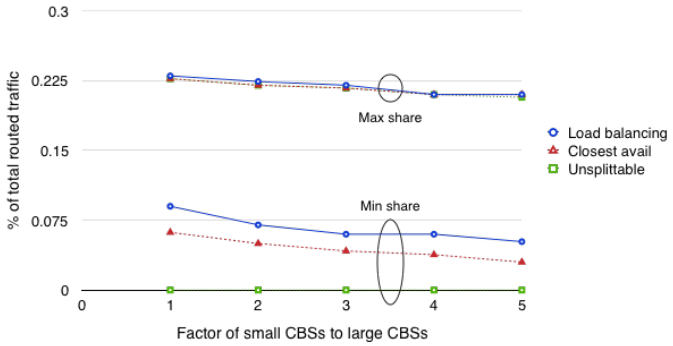

Figure 4.6: Minimum and maximum load share of a small (2nd tier) CBS depending on the ratio of small CBS number over large CBS number.

Fig. 4.6 shows (as in Fig. 4.4) the maximum and minimum traffic load routed to a small CBS depending on each policy. Even though for the FAIR policy more users are routed to the small CBSs overall, the maximum load share that one small CBS takes is almost the same for all policies. The increase in traffic load by the fair policy is distributed to the less loaded small CBSs. This is indicated by the higher minimum load among the stations. When applying either the CLOSEST AVAILABLE or the UNSPLITTABLE policy these CBSs are underused. Thus, the fair policy utilizes the available resources better.

4.6 Numerical Evaluations for Throughput Maximization

For evaluations for throughput maximization, 1000 cache-equipped networks were simulated on a square window of 8000 m × 8000 m of which the center 7000 m × 7000 m were evaluated to avoid edge effects. Users are distributed uniformly. The content placed into the CBSs comes from a 100 file catalog. The popularity of the content follows a Zipf distribution with parameter 0.6. In each simulation, the CBSs were placed following a PPP with density $0.8\frac{\text{CBS}}{\text{km}^2}$, resulting in 39.2 simulated CBSs with an average distance of 560 m to the closest CBS. Each CBS has a circular coverage radius of 1000 m and a power of $p = 1W$. The coverage zone is divided into an inner strong signal zone and a weak signal outer zone. The channel $h_{m,s}$ is calculated assuming a path loss exponent of 4. For the calculation of the SINR (see (4.7)), the distance of any user in the zone with strong signal ($\leqslant$ 500m distance to the CBS) is assumed to be 500m, whereas in the weak signal zone it is assumed to be 1000m. The noise is $\sigma^2 = 10^{-12}$W.

For each network, three different frequency reuse scenarios are simulated. In case of full frequency reuse ("1-freq"), each user uniformly gets assigned the bandwidth of $B = 1$MHz. All neighboring CBSs induce interference to each other. In a second scenario "2-freq", every station can operate on half of the total spectrum (random coin toss). To potentially serve the same number of users as in 1-freq, every associated user is served on half of the previous bandwidth, i.e. $B = 0.5$MHz. But in this case, the benefit is that only neighboring stations with the same half of the spectrum interfere. For "3-freq", the same principle is applied for a division of the spectrum into 3 orthogonal parts. For this run of experiments, the savings function is the linear weighted sum of traffic with the weights chosen as user throughputs (4.8).

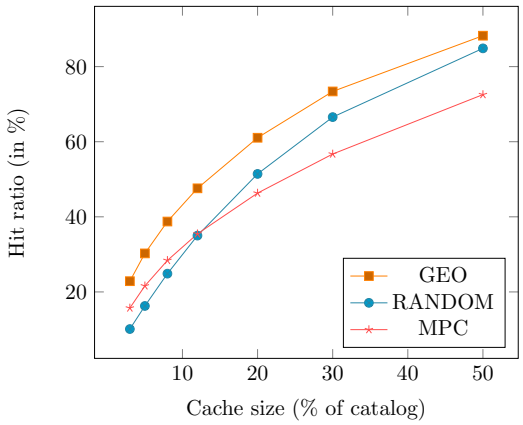

Figure 4.7: Hit ratio achieved for different content placements with user association maximizeing throughput under Opt- h and 1-freq.

The content is placed into the CBSs according to three different content placement policies known from the literature: globally most popular content (MPC), uniformly randomly placed content (RANDOM) with which each k-subset of files has equal probability to be cached, and geographic caching (GEO, see [BG15]) that places content in a probabilistic fashion taking coverage overlaps into account. User association is computed for each of these content placement policies with varying cache sizes, once maximizing throughput with Opt- h, once with conventional user association Closest.

Fig. 4.7 shows the hit ratio that can be achieved by Opt- h with the different content placement policies for the case 1-freq. For every content placement policy, the hit ratio increases with increasing cache size. The increases diminish when the cache size approaches 50% of the catalog. GEO is superior to the other policies in every case. For smaller cache sizes MPC is better than RANDOM, since the former provides the users with the most popular files. From a cache size of 12% of the catalog RANDOM performs better than MPC since more diverse content is offered to the users.

In Fig. 4.8, the relative gain of achieved hit ratio from Opt- h over Closest, i.e. the hit ratio achieved by Opt- h divided by the one achieved by Closest. In each instance RANDOM achieves the highest gain, followed by

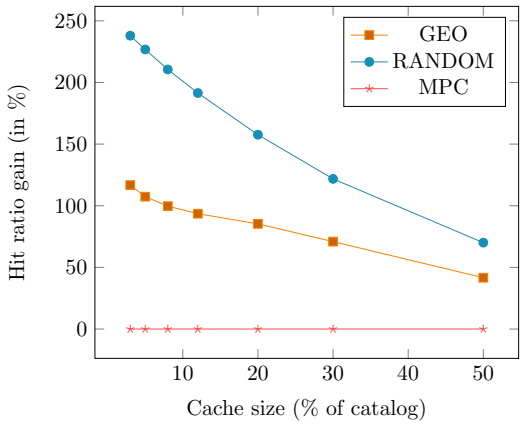

Figure 4.8: Relative Hit ratio gain with user association Opt- h maximizing throughput over Closest.

GEO. MPC does not achieve any gains at all. This shows that for RANDOM placement, Closest performs badly. This is due to the fact that RANDOM has a high probability of placing unpopular content into a CBS such that there are not many close users that contribute to the hit ratio, while some further away CBS might well cache the more highly requested content. With GEO, Opt- h still performs significantly better than Closest. GEO produces a diverse placement for the typical user, while still emphasizing more popular content at every CBS. This explains that the gain is between RANDOM and MPC. Since MPC places the same content at every CBS, it can never be the case that a station that has weaker signal stores content that a stronger one does not. Thus, Closest association is optimal in this case. The gains from Opt- h association decrease with increasing cache sizes, since more users find their requested content in the CBS that has the strongest signal. Overall, this plot shows that that user association policy makes a difference, particularly for content placement policies which emphasize content diversity amnong the CBSs.

Clearly, the gains in hit ratio from Opt- h over Closest come from the association of users with a weaker signal. In Fig. 4.9, we show that even if signal strength and interference are taken into account, Opt- h still is

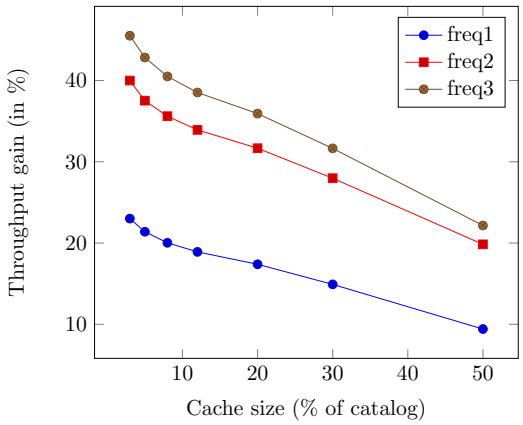

Figure 4.9: System throughput gained with user association maximizing throughput over CLOSEST.

superior to CLOSEST for GEO, the placement policy that achieved the best hit ratio. The metric used in this plot is the system throughput, the sum of users weighted by their achieved throughput $\sum_{m\in\mathcal{M}}\sum_{s\in\mathcal{S}}\sum_{f\in\mathcal{F}} w_{m,s,f}y_{m,s,f}$. The gain in thoughput from OPT- h over CLOSEST, i.e. the ratio of respective system thoughput, shows that even for freq-1, OPT- h achieves throughput gains over CLOSEST. These gains are even more pronounced for freq-2 and freq-3 networks which have reduced interference. When 3 frequency bands are assigned to the CBSs, a throughput gain of up to 45% can be achieved. While the gains diminish with increasing cache sizes, OPT- h provided a throughput gain of at least 10% in each simulated scenario, showing the significance of the user association under placement policies that diversify the content in neighboring CBSs.

4.7 Conclusions

This Chapter proposes an efficient distributed algorithm called Generalized Bucket-filling for cache-related user association. The algorithm can be applied for different purposes, for example hit ratio maximization, load-balancing and throughput maximization. Simulations show that the algo-

rithm effectively balances the user load, by distributing users from over- to underused CBSs. Furthermore, it is shown that under throughput maximization, the algorithm makes use of the possibility to associate CBSs to users that receive a weaker signal. This is particularly effective when orthogonal frequency bands are assigned to the CBSs.

Chapter 5

Cache Leasing and Content Placement

This chapter contains work from the published papers [KGD18] and [KGDR18]. Sections 5.2 and large parts of 5.4 are published in [KGD18]. The complexity analysis in Section 5.3 and Corollary3 come from [KGDR18]. The numerical evaluation (Section 5.2, in particular 5.6.3 to 5.6.6) are based on [KGD18] while the remaining evaluation in the subsections 5.6.7 to 5.6.9 come from [KGDR18]. Further analysis of the complexity of the problem SUR-T, in particular Corollary 1, is unpublished.

5.1 Introduction

Now, we change focus on the perspective of a CP. It is interested in placing its content closer to the user to improve QoS and to offload traffic from its central data storage infrastructure. An MNO offers to lease cache space at its CBSs for a certain time period for a fixed price per cache unit. Once the CP has decided how much memory to lease, it places its content, for example during an off-peak period. The cached content then remains static.

Assuming that an MNO fixes a price per cache unit at its CBSs, the

CP has two decisions to take: How much cache space to lease and what content to cache. As a basis for these decisions, we assume that the CP possesses fine grained spatio-temporal user traffic and content popularity data. The benefits, however, are only realized when users are associated to CBSs that cache the requested content. In order to predict which benefits can be generated from its cache leasing and content placement decisions, the CP needs to be informed about the MNO's *user association policy* (see previous chapter).

In this work, it is assumed that the CP's main objective is hit ratio maximization to offload traffic from its data storage. It will be shown that optimal leasing and placement decisions differ depending on the MNO association policy, CLOSEST or OPT- h. The OPT- h policy, however, allows for different secondary objectives. In the context of this work, these may be

- Load balancing among the CBSs: The CP may be interested to balance the *cache-related* load if it expects high traffic load to its cached content that threatens to overload the CBSs's wireless resources and thus diminish the QoS. The MNO might also incentivize the CP contractually not to overload any CBS by placing all its popular content in the same cache.
- Throughput maximization: In order to provide good QoS, the CP is interested in providing optimal network throughput for the cache-related users.

While focusing on the aforementioned CP aims, the following model is more general and is applicable to any convex, monotonuously increasing and continuously differentiable objective.

5.2 System Model and Problem Statement

5.2.1 Cache Leasing and Content Placement

We consider a cellular communications network with a finite set $\mathcal{M}$ of CBSs. Each CBS m is equipped with k_m memory units of size b_{MU} (in MBytes, e.g. 1000) which the CP can lease. Leasing and placement decisions are taken at the beginning of a long time window (and stay fixed throughout) during which content popularity statistics are assumed static. Denoting the decision variable of how many cache units to lease (see CP-1, Section 3.1.3) at m by $z_m \in \mathbb{Z}_{\geqslant 0}$, the bounded availability of memory gives the constraint set

$$z_m \leqslant k_m, \ \forall\, m \in \mathcal{M}. \tag{5.1}$$

The vector of the cache leasing variables is $\mathbf{z} = (z_m)_{m\in\mathcal{M}}$.

Having leased cache space at the CBSs, the CP places content from a finite object catalog $\mathcal{F}$ into the caches (see CP-2, Section 3.1.3). The decision to store content f in the cache of m will set the variable $x_{m,f}$ to 1, otherwise $x_{m,f} = 0$. The vector of content placement variables is $\mathbf{x} = (x_{m,f})_{m\in\mathcal{M},f\in\mathcal{F}}$. Each file has a given file size b_f (in MBytes), and all file-sizes are known. The limited capacity of the leased cache space gives the second constraint set

$$\sum_{f\in\mathcal{F}} b_f x_{m,f} \leqslant b_{\mathrm{MU}} z_m, \ \forall\, m \in \mathcal{M}. \tag{5.2}$$

For convenience, we define the set of feasible tuples of leasing and placement vectors as

$$\mathcal{X} \coloneqq \big\{(\mathbf{x},\mathbf{z}) \in \{0,1\}^{|\mathcal{M}||\mathcal{F}|} \times \mathbb{Z}_{\geqslant 0}^{|\mathcal{M}|} \;\big|\; (5.1),(5.2)\big\}.$$

5.2.2 Wireless Environment

For the wireless communication model, see Section 4.2.1. A summary of the notation:

- Every CBS m covers a set of regions $\mathcal{S}(m)$. The partition of the network area into regions $\mathcal{S}$ depends on the MNO's user association policy Π. The set of CBSs covering region s is denoted by $\mathcal{M}(s)$.
- For each region s, the CP has an estimate $N_{s,f}$ of the number of users requesting each file f. The vector of these popularity entries is denoted by $\mathbf{N}$. Note that the model does not assume spatially uniform traffic and that the vector $\mathbf{N}$ is general.

5.2.3 CP Savings and MNO Policy

The CP's aim is to maximize its financial savings from edge caching. These can, for example, be proportional to the hit ratio, the wireless throughput or take load-balancing of cache-related traffic into account. Formally, the CP savings are measured by the CP savings function $\mathrm{h}^{\mathrm{CP}}(\cdot)$ that takes the placement vector $\mathbf{x}$ as argument and maps it onto the CP savings (in €).

To achieve maximum benefits from content placement decisions, the CP depends on the MNO's user association policy. Assume MNO pursues the conventional Closest user association. Then, each CBS is related to a unique region. Since the MNO associates all users in a region to the corresponding CBS, the CP can use its knowledge about user statistics in each region to place content according to its own criteria. If, however, the MNO applies a cache-aware user association policy such as Opt-h, the CP needs to take the MNO reactions to its own placement decisions into account. Particularly, the MNO has its own performance metric that is expressed in the savings function $\mathrm{h}^{\mathrm{MNO}}(\cdot)$. The MNO maximizes this savings function performing user association.

The function $\mathrm{h}^{\mathrm{MNO}}(\cdot)$ measures one of the listed performance criteria for the MNO, e.g. throughput. Then, for a given content placement $\mathbf{x}$ the optimal association vecor is $\mathbf{y}^{\mathrm{Opt-h}^{\mathrm{MNO}}}(\mathbf{x})$ as defined in Section 4.2.2. In this work, we assume that in every instace the savings functions h^{CP} and $\mathrm{h}^{\mathrm{MNO}}$

of the CP and the MNO, respectively, are UA-equivalent by the following definition: If for all $(\mathbf{x}, \mathbf{z}), (\mathbf{x}', \mathbf{z}') \in \mathcal{X}$:

$$\mathrm{h}^{\mathrm{CP}}(\mathbf{x}) \leqslant \mathrm{h}^{\mathrm{CP}}(\mathbf{x}') \iff \mathrm{h}^{\mathrm{MNO}}(\mathbf{y}^{\mathrm{Opt-h}^{\mathrm{MNO}}}(\mathbf{x})) \leqslant \mathrm{h}^{\mathrm{MNO}}(\mathbf{y}^{\mathrm{Opt-h}^{\mathrm{MNO}}}(\mathbf{x}')),$$

then h^{CP} and $\mathrm{h}^{\mathrm{MNO}}$ are *UA-equivalent* or $\mathrm{h}^{\mathrm{CP}} \sim \mathrm{h}^{\mathrm{MNO}}$. This can be justified by the leasing contract between CP and MNO, and also by the common interest in caching. Non-UA-equivalent savings functions are subject for future research.

In Section 4.3, it was shown how performance criteria for user association can be translated into MNO savings functions. The above definition allows the translation of performance criteria for content placement into CP savings functions: A CP savings function h^{CP} measures a performance criterion, if an UA-equivalent MNO savings function $\mathrm{h}^{\mathrm{MNO}}$ exists that measures the performance criterion.

Now, additionally to the previous assumptions, we assume that there is a convex, increasing and differentiable relationship g that maps the MNO savings onto the CP savings, i.e.

$$\mathrm{h}^{\mathrm{CP}}(\mathbf{x}) = g \circ \mathrm{h}^{\mathrm{MNO}}(\mathbf{y}^{\mathrm{Opt-h}^{\mathrm{MNO}}}(\mathbf{x})). \tag{5.3}$$

Then, let

$$\mathrm{h} = g \circ \mathrm{h}^{\mathrm{MNO}}. \tag{5.4}$$

Note that h as the composition two increasing, continously differentiable, convex functions is itself increasing, continously differentiable and convex.

Three performance criteria for content placement are in the focus of this chapter:

HR) In case that the CP is solely interested in maximizing the hit ratio, the corresponding MNO savings function $\mathrm{h}^{\mathrm{MNO}}$ can be chosen as a *linear* function.

LB) If the CP maximizes the wireless throughput, the UA-equivalent function $\mathrm{h}^{\mathrm{MNO}}$ is *weighted linear*.

TP) The CP can include aspects such as soft resource requirements and load-balancing when $\mathrm{h}^{\mathrm{MNO}}$ is the sum of *strictly concave* functions (one function per CBS).

5.2.4 Problem statement

The objective of the CP is to lease cache memory at the CBSs and place content into it such that the relation of its expected savings to the leasing cost is optimal. As mentioned, the savings are given by the function $\mathrm{h}^{\mathrm{CP}}(\cdot)$ that takes as input the content placement action $\mathbf{x}$. The leasing costs at each CBS m are the product of leased units z_m times the price per unit p_m that is set by the MNO. Through this price, the CP is charged for making use of cache memory. An additional fee for the appropriate user association and content delivery can be included. Formally, the CP seeks a feasible tuple of vectors $(\mathbf{x}, \mathbf{z}) \in \mathcal{X}$ that maximizes the objective function

$$\begin{aligned} &\mathrm{h}^{\mathrm{CP}}(\mathbf{x}) - \sum_{m \in \mathcal{M}} p_m z_m \\ \overset{(5.3)}{=}\ & g \circ \mathrm{h}^{\mathrm{MNO}}(\mathbf{y}^{\Pi}(\mathbf{x})) - \sum_{m \in \mathcal{M}} p_m z_m \\ \overset{(5.4)}{=}\ & \mathrm{h}(\mathbf{y}^{\Pi}(\mathbf{x})) - \sum_{m \in \mathcal{M}} p_m z_m \end{aligned}$$

The CP's Cache Leasing and Content Placement problem (CLCP) can be formulated as the Non-Linear Mixed-Integer Problem (NLMIP)

$$\begin{aligned} (\mathrm{CLCP}) \max_{\substack{(\mathbf{x},\mathbf{z}) \in \mathcal{X} \\ \mathbf{y} \in \mathcal{Y}^{\Pi}}} \quad & \mathrm{h}(\mathbf{y}) - \sum_{m \in \mathcal{M}} p_m z_m \\ \text{s. t.} \quad & y_{m,s,f} \leqslant N_{s,f} x_{m,f}, \ \forall\, m, s, f, \end{aligned}$$

where m is a CBS, s is a planar network region and f is a data file.

Special Case: For zero costs and unit file size under the CLOSEST policy with linear savings function h, it is optimal to store k_m locally Most Popular Content in each CBS m.

5.3 Complexity

Even with a linear savings function h and without taking cache leasing into account, the CLCP problem is NP-hard.

Proposition 2. *The CLCP problem is NP-hard.*

Proof. Assuming that h can be evaluated in polynomial time, a certificate can be checked in polynomial time, thus the problem is in NP.

Now, we show a polynomial-time reduction from the Helper Decision problem (HD) presented in [SGD+13]. The reduction identifies users (in HD) with regions (in CLCP) and helpers (HD) with CBSs (CLCP). An instance of HD is transformed into a CLCP instance in the following way: Setting $b_f = b_{\mathrm{MU}} = 1$ and $k_m = M$ and all prices p_m to 0 eliminates the variables z_m and provides that (5.1) and (5.2) are equivalent to the capacity constraint in HD.

Choose $\mathrm{h}(\mathbf{y}) = \sum_{s\in\mathcal{S}} \tilde{w}_s \sum_{f\in\mathcal{F}} \sum_{m\in\mathcal{M}(s)} y_{m,s,f}$ where the $\tilde{w}_s$ correspond to the weights in HD. Since $\tilde{w}_s > 0$ for all s by definition in HD, and since $y_{m,s,f} \leqslant N_{s,f} x_{m,f}$ as well as (4.2) for all m, s, f, then whenever a CBS covering a region s stores the requested content f, all user traffic from the region will be associated with some CBS. Thus, $\sum_{m\in\mathcal{M}(s)} y_{m,s,f} = N_{s,f}$ if $x_{m,f} = 1$ for some $m \in \mathcal{M}(s)$ and 0 otherwise. Choosing $N_{s,f}$ (CLCP) as P_f (HD) for all regions s shows that, defined this way, the objective function of CLCP is equivalent to the objective function of HD.

This is a polynomial time transformation which concludes the proof. Note that in [SGD+13] NP-completeness is proved for catalog size two. □

5.4 Solution

The CLCP problem is very difficult to be solved even numerically by existing software, due to its high complexity. It is a mixed-integer problem with non-linear objective. We thus need to proceed analytically. The solution technique that resolves this problem is Generalized Benders decomposition by Schrijver [Sch86] and Geoffrion [Geo72] which converges to the global optimum. This method can decompose our problem in a way that removes the non-linearity from the discrete problem. The decomposed discrete problem, called Master, is linear and thus simpler to deal with. As a second by-product there appears a continuous convex subproblem called Slave, which can be solved by standard techniques (e.g. Lagrangian). Observe that due to Proposition 2, our solution algorithm cannot be polynomial even with linear savings function unless P = NP. However, a state-of-the-art MIP solver can be used for the iterative solution of Master. Its performance is shown in Section 5.6. In what follows, we give an overview over Generalized Benders decomposition applied to CLCP.

CLCP can be decomposed into two problems called Master and Slave. Master decides about cache leasing and content placement in the prefetching phase. Slave computes the optimal user association for a fixed content placement in the delivery phase. We obtain

$$\text{(Master)} \qquad \max_{(\mathbf{x},\mathbf{z})\in\mathcal{X}} \quad \mathrm{h}(\mathbf{y}(\mathbf{x})) - \sum_{m\in\mathcal{M}} p_m z_m,$$

where $\mathrm{h}(\mathbf{y}(\mathbf{x}))$ is the objective value of

$$\text{(Slave)} \qquad \mathbf{y}(\mathbf{x}) = \arg\max_{\mathbf{y}\in\mathcal{Y}^{\Pi}} \quad \mathrm{h}(\mathbf{y}) \qquad \text{s. t.} \quad (4.2).$$

Note that the Master problem can be treated by the CP, the Slave by the MNO. $\mathcal{X}$ is discrete and finite and $\mathcal{Y}^{\Pi}$ is compact and convex. Slave is

the UA-Π problem from Section 5.2.3 which is generally non-linear. Thus, Master cannot be solved directly. It can be solved, however, by following an iterative procedure that deals with this problem by solving a sequence of Slave problems for different values of $\mathbf{x}$ (and $\mathbf{z}$). The solutions to Slave are used to construct (linear) Benders cuts that constitute approximations to Slave. The Benders cuts are iteratively introduced as constraints of the *Surrogate Problems* which are linear approximations to Master. With each iteration, the approximation improves until the optimal solution of Master is found.

5.4.1 Benders Cuts

Let $\{(\mathbf{x}^t, \mathbf{z}^t) \in \mathcal{X} \mid t = 1, \ldots, T\}$ be a set of vector tuples feasible to Master for some $T \geqslant 0$. Let $\mathbf{y}^t := \mathbf{y}(\mathbf{x}^t)$ denote a corresponding vector that optimizes Slave for given $\mathbf{x}^t$. Let $\boldsymbol{\lambda}^t = (\lambda^t_{m,s,f})_{m\in\mathcal{M},s\in\mathcal{S}^\Pi(m),f\in\mathcal{F}}$ be the vector of Lagrangian multipliers corresponding to the constraints (4.2).

Slave is a convex problem. Thus, the duality theorem of convex programming implies

$$\mathrm{h}(\mathbf{y}(\mathbf{x})) \leqslant \quad \mathrm{h}(\mathbf{y}^t) + \sum_{m\in\mathcal{M}} \sum_{s\in\mathcal{S}^\Pi(m)} \sum_{f\in\mathcal{F}} \lambda^t_{m,s,f}(N_{s,f}x_{m,f} - y^t_{m,s,f})$$

for all feasible vectors $\mathbf{x}$. This upper bound to Slave is called *Benders cut*. Reformulated, we get

$$\mathrm{h}(\mathbf{y}(\mathbf{x})) \leqslant \Gamma^t + (\mathbf{v}^t)'\mathbf{x}, \tag{5.5}$$

where $\Gamma^t := \mathrm{h}(\mathbf{y}^t) - \sum_{m\in\mathcal{M}} \sum_{s\in\mathcal{S}^\Pi(m)} \sum_{f\in\mathcal{F}} \lambda^t_{m,s,f} y^t_{m,s,f}$ and $(\mathbf{v}^t)'$ is the transpose of $\mathbf{v}^t = (v^t_{m,f})_{m\in\mathcal{M},f\in\mathcal{F}}$ with $v^t_{m,f} = \sum_{s\in\mathcal{S}^\Pi(m)} \lambda^t_{m,s,f} N_{s,f}$.

5.4.2 Surrogate Problem

For a set of T Benders cuts (for some $T \geqslant 1$), we obtain an upper bound to the original problem CLCP by solving the *Surrogate* IP

$$\begin{aligned} (\text{SUR-}T) \max_{\substack{(\mathbf{x},\mathbf{z})\in\mathcal{X} \\ \gamma\in\mathbb{R}_{\geqslant 0}}} \quad & \gamma - \sum_{m\in\mathcal{M}} p_m z_m \\ \text{s. t.} \quad & \gamma \leqslant \Gamma^t + (\mathbf{v}^t)'\mathbf{x} \qquad t = 1,\ldots,T. \end{aligned}$$

The optimal objective value to the surrogate problem is denoted by SUR^T. The optimal solution consists of $\mathbf{x}^{T+1}$, $\mathbf{z}^{T+1}$ and γ^{T+1}. Note that the auxiliary variable γ, together with the Benders cuts, approximates Slave linearly. This way, SUR-T avoids the non-linearity which creates the difficulty for solving Master.

5.4.3 Benders Iteration and Convergence

Generalized Benders decomposition is an iterative process. We start in step 0 with an initial feasible tuple of leasing and placement vectors $(\mathbf{x}^0, \mathbf{z}^0) \in \mathcal{X}$ and without any Benders cuts. At the start of step $T \geqslant 0$, we have the current leasing and placement $(\mathbf{x}^T, \mathbf{z}^T) \in \mathcal{X}$ and T Benders cuts.

Slave: We solve Slave with input $\mathbf{x}^T$ to obtain the optimal association vector $\mathbf{y}^T$. Clearly, since the triple $\mathbf{x}^T$, $\mathbf{z}^T$ and $\mathbf{y}^T$ are feasible to the original problem CLCP, its corresponding objective value provides a lower bound to the optimal value of CLCP. Additionally, we compute the Lagrangian multipliers $\boldsymbol{\lambda}^T$ and the corresponding $(T+1)$-th Benders cut (5.5).

Surrogate: With the Benders cut we obtain the surrogate MIP SUR-$(T+1)$. Its optimal solution is the feasible leasing and placement vectors $(\mathbf{x}^{T+1}, \mathbf{z}^{T+1}) \in \mathcal{X}$. The objective value SUR^{T+1} is an upper bound to CLCP.

This process iterates. At any step T, SUR^T is the current upper bound (note that $\text{SUR}^{T+1} \leqslant \text{SUR}^T$ after every step T), while the current lower bound is provided by the best found solution $\max_{t\in\{0,\ldots,T\}} \mathrm{h}(\mathbf{y}^t) - \sum_{m\in\mathcal{M}} p_m z_m^t$. The

process terminates in the globally optimal cache leasing and content placement vectors $\mathbf{z}^*$ and $\mathbf{x}^*$ when the upper and lower bounds coincide. Convergence is guaranteed from the proof of Theorem 2.4 in [Geo72] and the fact that the domain $\mathcal{X}$ of the Master is finite.

5.4.4 Complexity of the Surrogate Problem

In every step T, an instance of the Surrogate problem needs to be solved. In general, the Surrogate problem is not easily tractable as to the following proposition. It consists of a polynomial time reduction from the well-known Set Cover problem. Its minimization version is defined as follows:

Definition 1. *Let $\mathcal{T}$ be an index set, called* universe. *Let $\mathcal{F} \subseteq 2^{\mathcal{T}}$ be a family of subsets.* Set Cover *asks for a subset $\mathcal{S} \subseteq \mathcal{F}$ of minimum size such that $\bigcup_{\mathcal{S}} = \mathcal{T}$.*

Proposition 3. *For $T \in \mathbb{N}$ and general parameters $\Gamma^t \in \mathbb{R}$ and $\mathbf{v}^t \in \mathbb{R}^{|\mathcal{M}||\mathcal{F}|}$, $t = 1, \ldots, T$, $p_m \in \mathbb{R}$ for $m \in \mathcal{M}$ as well as domain $\mathcal{X}$, the Surrogate problem SUR-T is NP-hard.*

Proof. Objective function and constraints can be evaluated in polynomial time, thus the problem is in NP.

Now, we reduce Set Cover to SUR-T. In the Set Cover decision problem, the question is if there is a subset $\mathcal{S}$ of a collection of sets $\mathcal{F} \subseteq 2^{\mathcal{T}}$ with cardinality less than or equal to S such that $\bigcup_{\mathcal{S}} = \mathcal{T}$ where $\mathcal{T}$ is the universe . For the reduction, we identify the Benders cuts t with the elements in the universe $\mathcal{T}$. The files f are in bijection to the sets in the family $\mathcal{F}$. Note that then, the Set Cover decision problem can be written as: Are

there

$$\begin{aligned} \gamma \in \mathbb{R}, \mathbf{x} \in \{0,1\}^{|\mathcal{F}|} \quad & \text{with} \\ \gamma &\geqslant 1 \\ \gamma &\leqslant \sum_{f \in \mathcal{F}} \mathbb{1}_{t \in f} x_f \qquad \forall\, t \\ \sum_{f \in \mathcal{F}} x_f &\leqslant S, \end{aligned}$$

where $\mathbb{1}_{t \in f} = 1$ if $t \in f$ and 0 otherwise? Here, x_f is a binary variable taking value 1 if and only if f is element of $\mathcal{S}$ and γ represents the lowest frequency of any element of $\mathcal{T}$ in $\mathcal{S}$.

This is an instance of the SUR-T decision problem for only one CBS (omitting the index m): The price is chosen as $p_m = 0$. The memory capacity is chosen as $k_m = S$, the file sizes as well as memory are all unit size ($b_f = b_{\mathrm{MU}} = 1$). Then, the variable z_m can be omitted since in the optimum, $\sum_{f \in \mathcal{F}} x_f = z_m$. The constraint $\sum_{f \in \mathcal{F}} x_f \leqslant S$ is the combination of the constraints (5.1) and (5.2). Constraint set $\gamma \leqslant \sum_{f \in \mathcal{F}} \mathbb{1}_{t \in f} x_f$ comes from choosing $\Gamma^t = 0$ for all t and $v^t_{m,f} = \mathbb{1}_{t \in f}$.

The reduction is performed in polynomial time. This concludes the proof. □

Set Cover is a well known problem. Bellare et al [BGLR93] proved that, unless P = NP, Set Cover cannot be approximated with constant factor in polynomial time. Feige [Fei98] showed that, unless NP has slightly superpolynomial time algorithms, there is no better polynomial-time approximation than $\log T$ where T is the size of the universe $\mathcal{T}$. It is well known that the greedy algorithm reaches the logairthmic bound [Chv79].

The polynomial time reduction from Set Cover to SUR-T then implies that SUR-T cannot be approximated with a constant factor.

Corollary 1. *Unless P = NP, SUR-T cannot be approximated with constant factor in polynomial time.*

Despite the fact that SUR-T is NP-hard, state of the art MIP solvers such as CPLEX are capable of solving SUR-T in reasonable runtime.

5.4.5 Implementation Considerations

Here, we provide a short discussion about information exchange between the entities solving Slave and SUR-T towards the global optimum. The SUR-T problem that solves the leasing and content placement aspect requires information on the price per cache unit. For each Benders cut, the vector of Lagrangian multipliers (dual prices) of the user association together with the optimal objective value of Slave need to be communicated. Knowledge of the spatial popularity statistics in this phase is reasonably assumed.

On the other hand, the Slave problem as formulated needs to know the exact utility h of the CP together with a subset of the popularity statistics which is related to the content the CP aims to cache. Hence, although the content placement and user association decisions are treated separately in two intertwined optimization problems, Slave does require sensitive information from the CP.

To overcome the implicit conflict of interest, a first suggestion can be that the MNO informs the CP about its general association policy Π to cooperate or not, together with the cache price. The CP can then solve both problems and after having identified the optimal association vector it can give association suggestions to the MNO each time a potential cache-related user emerges. Obviously, such approach requires a more integrated interaction between the two actors. A second proposal could be that the MNO solves Slave by using estimates on the CP savings function and popularity data. The resulting association solution communicated to the CP via the Lagrangian multipliers will generate a different set of Benders cuts for the CP. Obviously, in this case the final solution is suboptimal. An interesting extension of the current research would be to mathematically investigate the quality of such

solution.

5.5 Distributed Solution Algorithm for Slave

A distributed solution algorithm for Slave is given in Chapter 4.

5.6 Experiments and Numerical Evaluation

5.6.1 Environment

We simulate cellular networks in an urban environment and calculate the optimal cache leasing and content placement for 6 cases which differ in savings function and user association policy: The savings function h is chosen as: HR) Linear as in (4.5) for hit ratio maximization. LB) The sum of utility functions as in (4.6) where all utility functions U_m are chosen as the natural logarithm to achieve proportional fairness for the user traffic at the CBSs, thus balancing the load among CBSs. TP) The weighted sum of utilities as in (4.3) where the weights are defined as in (4.8) and the utility functions are identities. This way, the sum of utilities is equal to the network throughput. For each of the three cases of savings functions, the MNO's user association policy is a) the MNO-CP cooperative policy Opt- h or b) the conventional policy Closest. In each case, we simulated 100 random sets of CBS positions as a Poisson Point Process. This means that their total number in each run is a random Poisson realization, and their positions are uniformly distributed in the simulation window. The density of the PPP is $0.8\frac{\text{CBS}}{\text{km}^2}$ for the cases HR) (linear savings) as well as TP) (weighted linear savings) and $0.6\frac{\text{CBS}}{\text{km}^2}$ for the cases LB) (log-savings). This implies an average minimal distance of 560m and 650m between the CBS positions, respectively. For the cases HR) and TP), the evaluation window has size $5000\times5000\text{m}^2$, while the cases LB) were evaluated in a $3000\times3000\text{m}^2$ window. The expected number

of CBSs in the evaluated windows is 20 for case HR) and TP) and 5.4 for case LB). In both cases, a larger area was simulated to avoid edge effects. The coverage radius varies from 400m to 1200m. The MNO price per unit size cache memory at all CBSs varies (0.01€-2.00€). The user population is distributed uniformly over the network with a density of 30 users per km^2. The total simulated file catalog contains 100 objects. The content popularity follows the Zipf distribution with parameter 0.6 unless explicitly stated otherwise. The available cache size from the MNO is set to the catalog size so that only the pricing influences cache leasing decisions.

5.6.2 Implementation

All simulations have been performed using a native JAVA simulation environment. User association corresponding to the solution of the slave problem in Sections 5.4 and Chapter 4 is entirely done by optimization algorithms developped in the context of this thesis. The surrogate problem SUR-T in Section 5.5 is solved using the state of the art mixed-integer problem solver IBM CPLEX 12.7.0 in combination with IBM ILOG CPLEX Optimization Studio. The experiments have been performed on a machine with a 2.40 GHz 16-core processor and 48 GB RAM.

5.6.3 Results for Hit Ratio Maximization

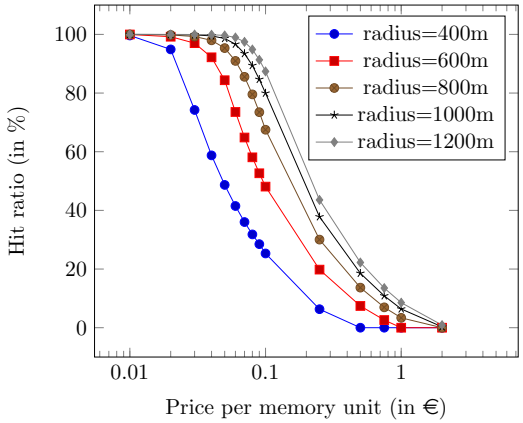

Figure 5.1: Hit ratio maximization: Hit ratio in relation to price for different coverage radii with Opt- h association.

At first, we present the simulation results in which the CP aims at hit ratio maximization (case HR)). On average, the optimal solution was obtained after 2 Benders iterations. We emphasize case HR.a) which performs Opt- h user association and compare it with case i.b) Closest. Fig. 5.1 illustrates how the hit ratio in case HR.a) depends on the price per cache unit for different coverage radii. For all radii, the hit ratio decreases with increasing prices. With lowest price (0.01€/Unit), the CP leases in each CBS the entire memory available, so the hit ratio is 100%. As the price increases, the CP leases less units, and the hit ratio is reduced. This happens more quickly in networks with smaller coverage areas because there are less users covered by each CBS and also less coverage overlap area. When the price reaches a high level (2€), the cost from leasing cache memory exceeds the benefit from cache hits and the hit ratio drops to 0% for all coverage radii. The differences between the curves in Fig. 5.1 diminish with higher radii where multi-coverage is already high enough.

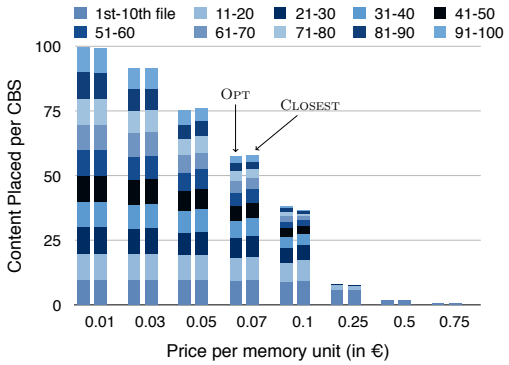

Figure 5.2: Hit ratio maximization: Cache lease and placement of popular files depending on cache unit price. Each left column represents the case of OPT- h association, each right column with CLOSEST association.

The CP's leasing and placement decisions for networks with coverage radius 1000m for the policies OPT- h (HR.a) and CLOSEST (HR.b) are shown in Fig. 5.2. For each price, there are two columns: The lefthand-side column represents the HR.a) case, the righthand-side column HR.b). The height of each column is the average amount of cache units per CBS which are leased for the respective price. The subdivisions of each column represent the popularity of the files stored in the leased memory space: The bottom part are the ten most popular files, the second-to-bottom part are the files of popularity rank 11 to 20 and so on. For the lowest cache price (0.01€), all 100 available units are leased: For both assignment policies, the amount of leased cache memory decreases with increasing price. For all prices, Fig. 5.2 shows that the less popular files are represented more frequently with OPT- h than with CLOSEST, especially for prices $\geqslant 0.1$€. There is more diversity of visible content with the OPT- h association.

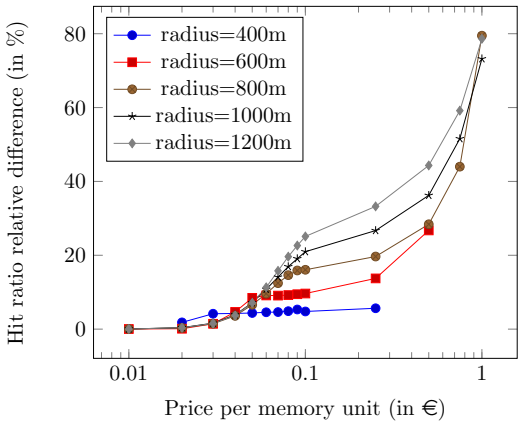

Figure 5.3: Linear savings function: Relative difference between the hit ratio achieved by optimal leasing and placement under Opt- h association to Closest association over cache unit price for different coverage radii.

Fig. 5.3 directly compares optimal CP decisions taken with Opt- h association with those taken with Closest association. For all prices the hit ratio achieved by Opt- h (case i.a) is higher than the one achieved by the Closest policy (case HR.b). The relative hit ratio differences are higher when the coverage area of the CBSs is higher. For higher prices, the Closest hit ratio is close to 0, therefore the relative differences can become very high.

5.6.4 Leasing Costs for Hit Ratio Maximization

While Figures 5.1 to 5.3 show the caching benefits for the CP, the costs it has to pay in return to the MNO (in case Opt- h) are depicted in Fig. 5.4. The CP costs equal the MNO income. This amount can be calculated by multiplying the number of leased cache units with the price per unit. The maximum of the curve can be clearly identified for each radius. *This is the operational point for the MNO when the latter aims for maximum income.* The maxima are higher for larger coverage areas, while the difference in income decreases with increasing radius. Furthermore, the higher the coverage radius, the higher the cache leasing price at which the maximum is achieved.

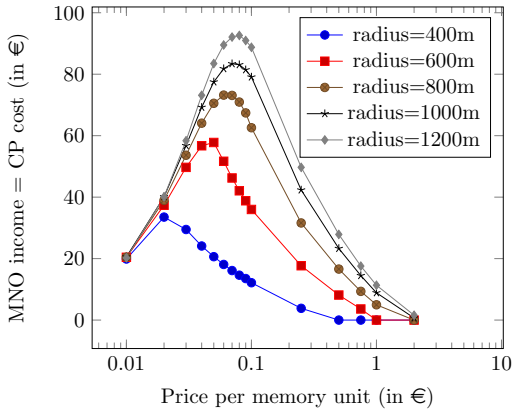

Figure 5.4: Hit ratio maximization: Income of MNO in relation to price per cache unit for different coverage radii with optimal leasing and placement under OPT- h association.

The relation between hit ratio and MNO income can be seen in Fig. 5.5: The x-axis displays the hit ratio achieved, the y-axis shows the income the MNO earns. Again, the income is higher in networks with larger coverage area. The maximum income for all simulated networks can be found for an achieved hit ratio between 80-90%. Conversely, if the CP decides to invest a certain sum, it can maximally achieve the rightmost of the two corresponding hit ratio values under the condition that the MNO chooses the pricing strategy most favorable to the CP.

5.6.5 Varying Content Popularity

The experimental results presented until now are based on a Zipf parameter of 0.6. However, a varying Zipf parameter influences the results: Fig. 5.6 shows that the higher the price, the lower the hit ratio for any Zipf parameter in case HR.a). This is due to the fact that lower price implies more leased units for the CP. For all prices (except the lowest one which achieves a hit ratio of near 100% throughout), the hit ratio increases with increasing Zipf parameter: With higher Zipf parameter, the population share requesting the most popular files becomes higher, thus caching popular files becomes

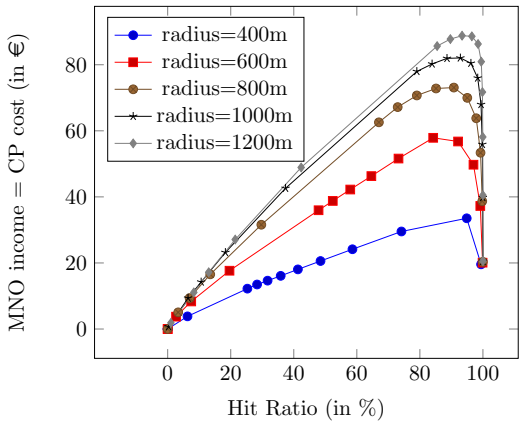

Figure 5.5: Hit ratio maximization: MNO income/CP investment over hit ratio for different coverage radii with Opt- h. Achieving the hit ratio on the x-axis results in the MNO income on y-axis.

more profitable. Also, for the same price, the leased cache memory is more effective with higher Zipf parameters. The lower the Zipf parameter, the more pronounced the differences in hit ratio between different cache prices since the benefits from overlapping coverage are bigger when content popularity is more even.

5.6.6 Results for Load-Balancing among CBSs

Here, we present the experimental results for load balancing (case LB)) in which the savings function h^{MNO} is the sum of logarithms. When such user association is performed, the CP aims at load balancing of cache-related traffic among the CBSs. The optimal solution was obtained after 8 Benders iterations on average. Fig. 5.7 shows the hit ratio for varying cache unit price both for the Opt- h (different coverage radii) and the Closest policies. Due to the specific choice of the logarithmic savings function of case LB) the Closest association gives identical cache leasing and content placement results for all coverage radii. For every coverage radius and every cache unit price, the Opt- h policy achieves a higher hit ratio than the Closest policy.

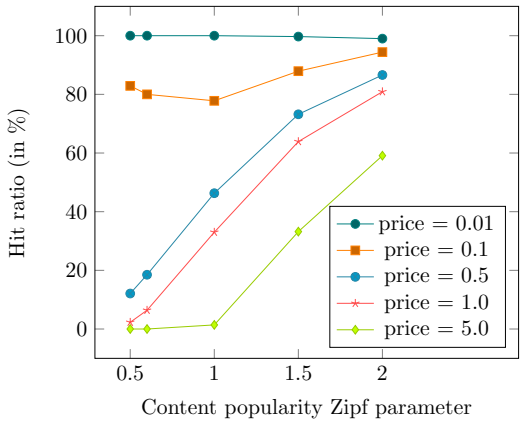

Figure 5.6: Hit ratio maximization: Hit ratio in relation to Zipf parameter for different prices with Opt- h.

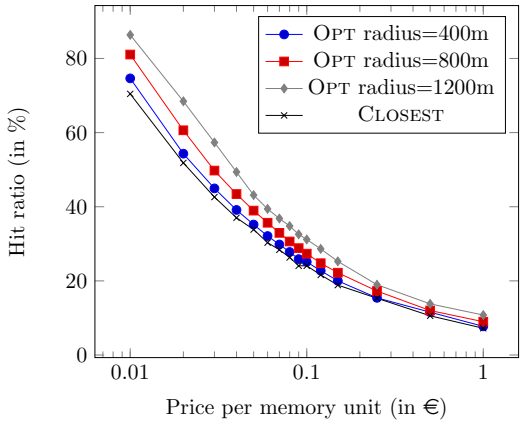

Figure 5.7: Load-balancing: Hit ratio in relation to price for different coverage radii for cases Opt- h and Closest.

Furthermore, the higher the coverage radius in case LB.a), the higher the hit ratio. The hit ratio improvement can reach over 15 percentage points using Opt- h. With increasing prices, caching becomes less profitable and the hit ratio decreases.

The advantage to the hit ratio of the Opt- h policy (LB.a) can be explained by the optimal content placement shown in Fig. 5.8. Each pair of columns represents cache leasing and content placement for a certain unit price. Each left column represents the optimal decisions with Opt- h associ-

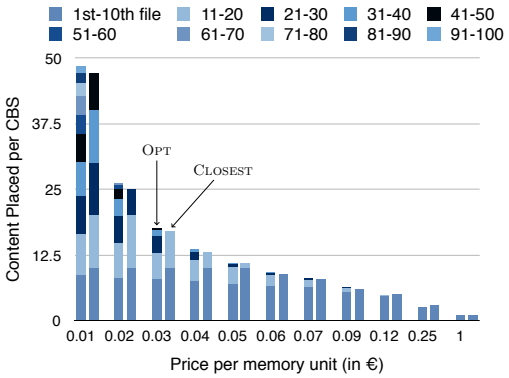

Figure 5.8: Load-balancing: Cache lease and placement of popular files depending on cache unit price. Each left column represents OPT- h, each right column is CLOSEST.

ation, each right column the optimal decisions taken for CLOSEST association. The height is the average amount of cache units leased per CBS. The inner sections of the columns represent the content placement in all of the CBSs: The lowest section are the 10 most popular files, the second lowest the files ranked 11 to 20 and so on. It can be seen that particularly for low cache unit prices, the diversity of cached content is higher in case LB.a) than in case LB.b). For the lowest price (0.01€), CLOSEST provides only content from the more popular half of the catalog, while the optimal content placement under OPT- h association places content from the tail of the catalog as well.

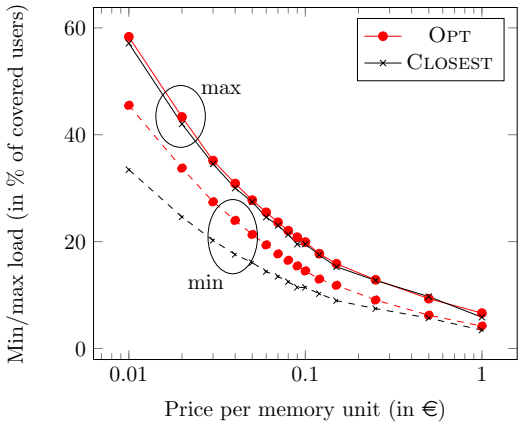

Figure 5.9: Load-balancing: Cache lease and placement of popular files depending on cache unit price. Each left column represents OPT- h, each right column is CLOSEST.

The main purpose of choosing the specific savings function in LB.a) is, however, the balancing of traffic load among the CBSs in order to avoid excess of resources by user overflow which will lead to service dissatisfaction. Fig. 5.9 shows that the additional load (from the increase in hit ratio using Opt- h, see Fig. 5.7), is distributed to the less loaded CBSs. The two upper (solid) lines in the graph represent the maximum load of a CBS in relation to the overall covered population per CBS both in the cases LB.a) and LB.b). The two lower (dashed) lines are the minium loaded CBS. The maximum loaded CBSs in both LB.a) and LB.b) coincide as the figure shows. The minimum loaded CBS of LB.a) is higher than the LB.b), showing that excess users coming from the higher hit ratio Figure 5.7) are associated to the less loaded stations.

The three plots show that the optimal leasing and placement under the Opt- h policy achieves an increase in hit ratio (good for both the CP and the MNO) while at the same time diversifying the cached content (good for the user) and avoiding an overload of CBSs (good for everybody).

5.6.7 Policy Comparison

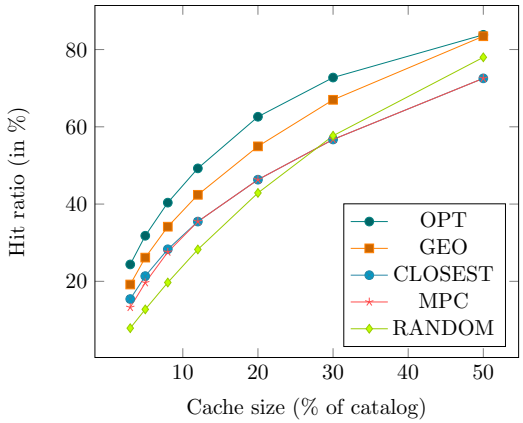

Figure 5.10: Hit ratio for different content placement policies with fixed cache sizes and spatially inhomogeneous traffic with linear savings function.

Note that in this chapter, until now, the cache leasing and content placement decisions that have been considered are solutions of the CLCP problem, depending on the MNO user association policy Opt- h or Closest. In this section, the leasing mechanism is not considered, and instead a varying limit for the cache size is imposed. Here, the Opt- h results are the results of the CLCP problem with 0 prices under linear savings function with Opt- h user association. The Closest results are the respective results with Closest user association. All other results are derived from different content placement policies together with Opt- h user association. Similar results have already been presented in Section 4.6, in which the optimized content placement policies are not yet included.

In order to illustrate the true benefits of Opt- h over Closest (with linear savings) as well as over other content placement policies, spatially inhomogeneous content popularity should be considered. We evaluate here a traffic scenario in which the network window is symmetrically divided such that on each side, the popularity of files follows a Zipf distribution with parameter 0.6, but the ten most popular files on one side are swapped with the second most popular decile on the other side to give locally differing popularity distributions. We fix the cache sizes uniformly and do not consider the leasing mechanism, enabling comparisons with known content placement policies that do not take cache leasing into account.

Fig. 5.10 shows that for a large range of cache sizes, Opt- h provides a 50% relative gain in hit ratio over Closest. Compared to other placement policies such as geographic caching (GEO, see [BG15]), caching of globally Most Popular Content (MPC) and uniformly random caching (RANDOM), the Opt- h policy provides considerable gains in each case. This result works strongly in favour of our suggestion that the MNO and CP should closely collaborate for user-association decisions.

5.6.8 Sensitivity to Traffic Estimation Errors

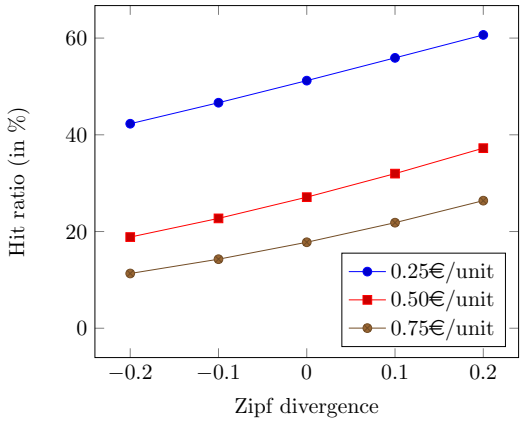

Figure 5.11: Hit ratio when the Zipf parameter of the actual content popularity diverges from the Zipf parameter on which the leasing and placement decisions are based.

The actual content request distribution can diverge from the statistics on which the cache leasing and content placement decisions are based. We evaluate the effect of a diverging Zipf parameter of the popularity distribution on the hit ratio of CP decisions based on parameter 0.6 (case HR)).

From Fig. 5.11 it can be seen that if the Zipf parameter was underestimated for leasing and placement, the hit ratio will be higher than expected as well. Conversely, an overestimated Zipf parameter leads to a decrease in the obtained hit ratio.

5.6.9 Results for Throughput Maximization

Finally, we present the simulation results for throughput maximization (case TP)). On average, the optimal solution was obtained after 2 Benders iterations. The CBSs have a coverage radius of 1000m and a power of $p = 1W$. Every station has two coverage zones, one in the inner half of the coverage radius, one in the outer half. The channel $h_{m,s}$ is calculated assuming a path loss exponent of 4. For the calculation of the SINR (see (4.7)), the distance of any user in the zone with strong signal ($\leqslant$ 500m distance to the CBS) is

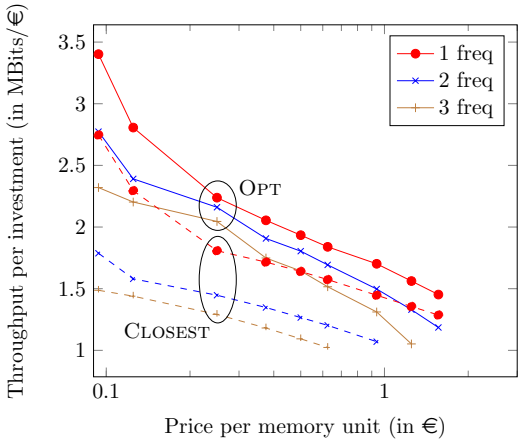

Figure 5.12: Throughput optimization by weighted savings function: Throughput per cache investment.

assumed to be 500m, in the weak signal zone it is assumed to be 1000m. The noise is $\sigma^2 = 10^{-12}$W.

For each network, three different frequency reuse scenarios are simulated. In case of full frequency reuse ("1-freq"), each user uniformly gets assigned the bandwidth of $B = 1$MHz. All neighboring CBSs induce interference to each other. In a second scenario "2-freq", every station can operate on half of the total spectrum (random coin toss). To potentially serve the same number of users as in 1-freq, every associated user is served on half of the previous bandwidth, i.e. $B = 0.5$MHz. But in this case, the benefit is that only neighboring stations with the same half of the spectrum interfere. For "3-freq", the same principle is applied for a division of the spectrum into 3 orthogonal parts.

For this run of experiments, the savings function is the linear weighted sum of traffic with the weights chosen as user throughputs (4.9). The valuation of the savings is set to $c = 1$€/MBits which implies that any achieved throughput of more than 1MBits per invested Euro brings the CP into profit. Fig. 5.12 shows the total throughput per invested Euro ("caching efficiency") over cache leasing price (between 0.1€ and 1.6€) for the user association

policies Opt- h and Closest, each for the different frequency reuse scenarios. Observe that Opt- h yields a significantly higher caching efficiency than Closest throughout. The relative difference between the two user association policies is greatest in the 2-freq scenario whereas the absolute system performance is maximised with 1-freq. This is due to the fact that for 2-freq, center users receive less bandwidth than in 1-freq even though interference is low. Interference reduction and system performance of 2-freq and 3-freq can be improved, however, by assigning the parts of the spectrum to CBSs in a more intelligent way than uniformly randomly. Depending on frequency reuse, the MNO can offer profit through caching ($\geqslant$ 1MBits/€) to the CP even when the cache-price is set high.

5.7 Conclusions

Extensive experiments for random network topologies allow to compare the optimal CP decisions for different MNO association policies, cache prices, as well as CP savings functions. In all versions of the problem, we have identified a unique price that maximizes the MNO revenue. It depends on how much the CP valorizes traffic offloading achieved by the edge caches. This information is included in the CP's choice of the savings function. Another main conclusion is that MNO association policies which adhere to CP actions and exploit multi-coverage opportunities achieve higher offloading benefits for a given monetary investment. All these results suggest that the CP and MNO can jointly develop cooperative business models related to caching, that lead to considerable economic as well as operational benefits for both parties.

Chapter 6

Competition between Multiple CPs

This chapter contains unpublished research and premliminary results.

6.1 Introduction

In the previous chapter, a CP leases edge memory space from an MNO for a fixed price. Here, we look at a scenario in which several CPs compete for the same cache resources of one MNO.

Each of the CPs has its own separate content catalog, disjoint from the others. The CPs all deliver their content via the MNO's wireless network that is equipped with caches. The cache memory is allocated to the CPs by the MNO in an allocation phase that precedes the content placement and delivery phases from the previous chapters.

Assume, for example, that the MNO to allocates cache space to each CP separately for a fixed price. In that case, the analysis for cache leasing in Chapter 5 applies. The CPs may, however, have varying numbers of users overall or locally, different shapes of popularity profiles, or might value caching effects differently. It is thus possible that, if the MNO offers a fixed

amount of cache space to each CP, one or several CPs decide not to lease the entirety of their allocated cache space while other CPs would like to lease beyond what they have been allocated. A higher aggregated hit ratio could then be achieved by allocating unused cache space to CPs that are willing to pay for the additional cache units. From the MNO point of view, it is preferable to utilize the installed caches in their entirety since more leased cache space induces the offloading of more network traffic. In economic turns, this means that the MNO aims at *clear the market*. The question arises how much cache space should be allocated to each CP at each CBS.

This chapter follows the idea of *social welfare maximization*: Every CP values a profile of allocated cache space through the savings that it can achieve. The social welfare is the sum of all CP savings. The aim of the MNO is to allocate its cache space to the MNOs that maximizes social welfare. The underlying assumption is that CPs that achieve higher savings help the MNO more since they are able to serve more users from the caches.

Once the CPs have cache space assigned, they will place their content into the caches in the content placement phase. The placement decisions are taken based on knowledge of their respective traffic and popularity data and the announced MNO user association policy. In the content delivery phase, cache hits generate savings to the CPs through the offloading of user traffic from their own data center infrastructure and through improvement of user QoS. The CPs are assumed to be rational entities that aim at maximizing their individual savings that they can achieve through such an edge caching scheme.

The CP savings, expressed through savings functions that map assigned cache space onto a financial value, are not necessarily known to the MNO. The communication between MNO and CPs instead occurs through a financial transaction. This can take place through an allocation mechanism in which the CPs are *price takers*, or through a game in which the CPs *antici-*

pate the bidding behaviour of their competitors.

6.2 System Model

6.2.1 Cache-equipped Network

Consider a cellular communications network with a finite set $\mathcal{M}$ of CBSs, operated by an MNO. Each CBS m is equipped with k_m memory units. A set of CPs $\mathcal{R}$ competes for access to the cache space. The amount of memory units that CP $r \in \mathcal{R}$ gets allocated at CBS $m \in \mathcal{M}$ is denoted by $z_m^{(r)}$. The allocation vector for r is $\mathbf{z}^{(r)} = (z_m^{(r)}), m \in \mathcal{M}$. At each CBS, the MNO can assign no more cache units than are installed, which implies that the following constraint holds:

$$\sum_{r \in \mathcal{R}} z_m^{(r)} \leqslant k_m, \ \forall\, m \in \mathcal{M}. \tag{6.1}$$

After CP r gets assigned memory space $\mathbf{z}^{(r)}$, it can place content from its catalog $\mathcal{F}^{(r)}$ in it.

6.2.2 CP Savings

Through cache hits, the CP generates savings that are expressed through the savings function $\mathrm{h}^{(r)} : \times_{m \in \mathcal{M}} \{0, \ldots, k_m\} \to \mathbb{R}$ that takes the cache allocation vector $\mathbf{z}^{(r)}$ as input and maps it onto the CPs savings. $\mathrm{h}^{(r)}$ is componentwise non-decreasing, i.e. there are never more savings generated to a CP by have less cache space allocated at a CBS, all others being equal.

The savings depend on the content that the CP places into the caches, and subsequently on the association of users to the caches. In Section 5.2.4, a function h^{CP} (here specific for r $\mathrm{h}^{\mathrm{CP}\text{-}(r)}$) is defined that takes the content placement vector $\mathbf{x} \in \{0,1\}^{|\mathcal{M}||\mathcal{F}|}$ as input. Since CP r behaves rationally and places content optimally, the savings from the assigned cache space $\mathbf{z}^{(r)}$ are

the solution of the problem

$$\mathrm{h}^{(r)}(\mathbf{z}^{(r)}) = \max_{\mathbf{x}\in\{0,1\}^{|\mathcal{M}||\mathcal{F}|}} \mathrm{h}^{\text{CP-}(r)}(\mathbf{x})$$
$$\text{s.t.} \sum_{f\in\mathcal{F}^{(r)}} b_f x_{m,f} \leqslant b_{\mathrm{MU}} z_m, \ \forall\, m \in \mathcal{M},$$

where b_f is the file size of file f and b_{MU} is the size of a leased memory unit.

The above problem is a 0-1 knapsack problem with general componentwise non-decreasing objective. It is connected to the CLCP from Section 5.2.4: If all prices in a CLCP are set to $p_m = 0$ and the leased cache units are fixed to the vector $\mathbf{z}^{(r)}$, then the savings $\mathrm{h}^{(r)}(\mathbf{z}^{(r)})$ can be calculated with the methods from Chapter 5.

6.2.3 Problem Statement

The assignment of cache space by the MNO to the CPs that maximizes social welfare is the solution to the following problem:

$$\text{(SW)} \qquad \max_{\mathbf{z}^{(r)}\in\mathbb{Z}_{\geqslant 0}^{|\mathcal{R}|}, r\in\mathcal{R}} \quad \sum_{r\in\mathcal{R}} \mathrm{h}^{(r)}(\mathbf{z}^{(r)})$$
$$\text{s.t.} \qquad (6.1).$$

This chapter provides a preliminary investigation of a relaxed and simplified version of SW, making three assumptions:

a) SW is relaxed by removing the integrality constraint on the variables $z_m^{(r)}$, as well as, implicitly, the variables $x_{m,f}^{(r)}$. Thus, we allow assignment of fractional cache space. Fractional content placement can be interpreted as probabilistic content placement.

b) SW is simplified by imposing that all CBSs have the same amount of cache space k installed.

c) The cache memory is split in the same way at every CBS, i.e. $z_m^{(r)} = z_{m'}^{(r)}$ for all m, m'. This uniform cache allocation to CP r is denoted by $z^{(r)}$.

Then, the simplified and relaxed problem is

$$\text{(SWsr)} \qquad \max_{z^{(r)} \in \mathbb{R}_{\geqslant 0}^{|\mathcal{R}|}, r \in \mathcal{R}} \quad \sum_{r \in \mathcal{R}} \mathrm{h}^{(r)}(z^{(r)}) \qquad \text{s.t.} \quad \sum_{r \in \mathcal{R}} z^{(r)} \leqslant k.$$

Note that, under the relaxation of the integrality constraints and the definition of the savings function $\mathrm{h}^{(r)}$ is reduced to

$$\mathrm{h}^{(r)}(z^{(r)}) = \max_{\mathbf{x} \in [0,1]^{|\mathcal{M}||\mathcal{F}|}} \mathrm{h}^{\text{CP-}(r)}(\mathbf{x}) \quad \text{s.t.} \sum_{f \in \mathcal{F}^{(r)}} b_f x_{m,f} \leqslant b_{\mathrm{MU}} z^{(r)}, \ \forall\, m \in \mathcal{M}.$$

By the definition of $\mathrm{h}^{\text{CP-}(r)}$, $\mathrm{h}^{(r)}$ is concave, non-decreasing and continuous.

6.3 Cache Allocation Mechanisms

6.3.1 Proportional Allocation

Since the savings functions $\mathrm{h}^{(r)}$ are not available to the MNO, the problem SWsr is not solved in a straightforward way but through the *proportional allocation mechanism* [NRTV07, Chapter 21]. Each CP r places a bid d_r to the MNO for cache access. The bid d_r represents the financial investment that CP r is willing to make to gain access to the caches. Then, the MNO allocates the cache space to the CPs such that the values $z^{(r)}$ are proportional to d_r. This is achieved by making all CPs pay the same price p per memory unit such that $z^{(r)} = d_r/p$. The entire cache memory is used, i.e. the market is cleared, if the price p is chosen such that

$$\sum_{r \in \mathcal{R}} \frac{d_r}{k} = p. \tag{6.2}$$

This is possible if $\sum_{r \in \mathcal{R}} d_r > 0$.

6.3.2 CPs as Price Takers

A CP r acts as a price taker if, given a price per memory unit, the resulting bid d_r does not take potential future price changes into account. Given a price p, a price taking CP aims at maximizing its payoff function

$$P_p^{(r)}(d_r) = \mathrm{h}^{(r)}(d_r/p) - d_r. \tag{6.3}$$

If for some vector $\mathbf{d} = (d_r), r \in \mathcal{R}$ and a price p, it is true that the respective payoff function (6.3) is maximized and the equality (6.2) holds, then $\mathbf{d}$ is a *competetive equilibrium*. In [NRTV07, Chapter 21], it is shown that any competetive equilibrium is optimal for the problem SWsr.

This equilibrium can be achieved iteratively, letting the MNO adjust the price p step by step.

6.3.3 Price Anticipating CPs

In the previous section, CPs choose their bids maximizing their payoff function (6.3), treating the price p as a fixed parameter. Here, the CPs are aware that the price is set according to (6.2), i.e. that the share of the total memory k that CP receives depends on the other bids by the ratio $d_r / \sum_{r'} d_{r'}$. Knowing this, the CPs adjust their playoff accordingly. This makes the process a game between the CPs with the payoffs

$$Q^{(r)}(d_r; \mathbf{d}_{-r}) = \begin{cases} \mathrm{h}^{(r)}(k \cdot d_r / \sum_{r'} d_{r'}) - d_r & \text{if } d_r > 0, \\ \mathrm{h}^{(r)}(0) & \text{otherwise,} \end{cases}$$

where $\mathbf{d}_{-r}$ is the vector of all values $d_{r'}$ for $r' \neq r$.

A *Nash equilibrium* of the game defined by $Q_r, r \in \mathcal{R}$ is a vector $\mathbf{d}$ such that for all r:

$$Q_r(d_r; \mathbf{d}_{-r}) \geqslant Q_r(\tilde{d}_r; \mathbf{d}_{-r}) \quad \text{for all } \tilde{d}_r \geqslant 0.$$

In [NRTV07, Chapter 21] it is shown that such a Nash equilibrium exists with $\sum_r d^{(r)} > 0$ and that it is unique. The related vector $(z^{(r)}), r \in \mathcal{R}$ is the solution to the following optimization problem:

$$\begin{aligned} \text{(GAME)} \qquad & \max_{z^{(r)} \geqslant 0, r \in \mathcal{R}} && \sum_{r \in \mathcal{R}} \hat{\mathrm{h}}^{(r)}(z^{(r)}) \\ & \text{s.t.} && \sum_{r \in \mathcal{R}} z^{(r)} \leqslant k. \end{aligned}$$

where

$$\hat{\mathrm{h}}^{(r)}(z) = (1 - z/k)\, \mathrm{h}^{(r)}(z) + (z/k)\Big(1/z \int_0^z \mathrm{h}^{(r)}(t)dt\Big).$$

In [NRTV07, Chapter 21], the author shows that the Nash equilibrium loses no more than 25% in comparison to the optimal solution of SWsr, i.e. if $\mathbf{z}_{\mathrm{G}} = (z_{\mathrm{G}}^{(r)})$ is a solution to GAME and $\mathbf{z}_* = (z_*^{(r)})$ is optimal to SWsr, then

$$\sum_{r \in \mathcal{R}} \mathrm{h}^{(r)}(z_{\mathrm{G}}^{(r)}) \geqslant \frac{3}{4} \sum_{r \in \mathcal{R}} \mathrm{h}^{(r)}(z_*^{(r)}).$$

6.4 Conclusions

This section models multiple CPs that compete for the same cache resources. The MNO acts as a neutral arbiter that sets cache prices such that the cache resources are completely used. A game is designed that, for the case that each cache is divided in the same way among the CPs and relaxing the integrality of cache units, provides a high quality Nash equilibrium. An optimization problem is given that helps attaining the Nash Equilibrium in practice.

The following questions, relating to the assumptions made in Section 6.2.3, arise from the preliminary results:

a) How can the problem be solved if each cache is seen as a separate resource, allowing different cache allocations at each CBS?

b) Does the game change when each CBS is equipped with memory of different size k_m?

c) How can the approach be extended to discrete cache partitioning?

Chapter 7

Conclusions and Future Work

7.1 Conclusions

This thesis develops a business model for wireless edge caching. Mobile network operators (MNOs) physically install and maintain caches at cached base stations. Cache space is leased to content providers (CPs). The CPs place their content into the caches for a time period during which the cache content remains stable. The placement decisions are based on CP user traffic and content popularity data as well the MNO user association policy. For users in coverage overlap areas, the user association policy decides which CBSs users will be associated to. The association policies can be cache-aware, i.e. taking the cached content into account, or cache-oblivious.

In a first step in Chapter 4, MNO cache-aware user association is investigated. An efficient algorithm called Generalized Bucket-filling is developped for the calculation of cache-aware user association optimizing different performance criteria. The main takeaways from its application are:

- Cache-aware user association yields big advantages over conventional cache-oblivious user association for a given content placement. In particular, performance criteria such as the hit ratio or system throughput

can be improved significantly.

- User association performance is improved particularly if neighboring CBSs store different content.
- When load balacing is the objective, a high hit ratio can be achieved without overloading CBSs but shifting additional traffic to less used stations.
- The best user association performance is achieved when the content is placed preempting user association.
- Cache-aware user association performs better when neighboring stations use different frequency bands, reducing the influence of interference

In a second step in Chapter 5, a scenario in which one MNO leases cache space to one CP was examined. The problem of optimal cache leasing and content placement under cache-aware user association is decomposed into an MNO and a CP part. The resulting leasing and placement policy that optimizes CLCP when Opt- h user association is applied is (in slight abuse of notation) also called Opt- h. The related experiments mainly show the superiority of Opt- h over the optimal cache leasing and content placement decisions under the cache-oblivious association policy Closest. The takeaways are:

- The larger coverage overlap area of CBSs in the network, and the higher the cache lease price, the better Opt- h performs in relation to Closest.
- Opt- h cache leasing and content placement provides a more diverse content cached catalog than Closest.
- The MNO can identify the operational point for pricing that maximizes its income. The optimal price depends on the CBS coverage radius.
- The effectiveness of edge caching depends on the content popularity statistics. The more skewed the statistics are towards fewer popular

files, the more effective is edge caching.

- While the proposed edge caching scheme depends on content popularity statistics, smaller inexactitudes in them only have minor effects on the scheme's effectiveness.
- Leasing and placement based on Opt-h is superior to Closest particularly in scenarios in which inter-cell interference is reduced through the use of orthogonal frequency bands on neighboring stations.

Beyond the results from Chapters 4 and 5, Chapter 6 shows a way towards evaluating scenarios in which several CPs compete for the cache resources of one MNO. In such a scenario, the CPs can act as price takers, accepting the leasing price given out by the MNO, or they can be price anticipating. In the latter case, the CPs preempt the behavior of the competing CPs, assuming that the MNO is interested in leasing out the entire available cache space. In case the CPs are price anticipating, any equilibrium of the resulting game loses no more than 25% of efficiency against the social optimum.

7.2 Future Work

There are several ways in which these results can be extended.

7.2.1 Data Uncertainty

User association in this work is treated in an accumulative sense, based on user traffic and content popularity predictions. The resulting distribution of cache-related users is calculated, not the association of individual users. While simulations show that slight inexactitudes in these predictions do not affect the effectiveness of the proposed edge caching scheme, further questions can be asked:

- How can knowledge about data uncertainty be included in the model? For example, it is possible that traffic and popularity predictions for

certain content is more exact than for other content, or the predictions differ depending on location. Statistics on data exactness can be taken into account applying methods of stochastic optimization.

- How should a mechanism be designed that associates individual users to the CBSs in the delivery phase, such that the distribution of users closely matches the results of the accumulative user association in Chapter 4?

7.2.2 Non-cooperation between MNO and CP

In Chapter 5, a CP optimizes cache leasing and content placement decisions, given the MNO user association policy. It is assumed that the objectives of CP and MNO regarding user association are equivalent, i.e. that the MNO associates users in a way that maximizes the CP savings. The work can be continued by investigating:

- How are CP decisions affected if the MNO associates users aiming at a different objective than the MNO? For example, the CP might want to maximize throughput while the MNO aims at load balancing.

7.2.3 Competing CPs and competing MNOs

Chapter 6 shows a way towards a model for an edge cache market. The steps further in that direction include:

- The model in Section 6.3 allocates the cache space to several CPs at a generic CBS. How can the more general model be solved when different cache allocations are allowed at each CBS?
- In the solution to the cache allocation model, the integrality constraint on leasing and placement are relaxed. How can the model be solved including integrality, allocating entire cache units?
- The model can furthermore be extended to include several competing

MNOs that offer edge caching services to the CPs. The MNOs that potentially have different user demographics which implies different content popularity statistics per MNO for a CP. The MNOs might also offer different user association policies. Each CP then needs to decide how to split the financial resources among the MNOs to maximize their savings.

7.2.4 Deeper or Interconnected Caches

The usefulness of edge caching has been shown throughout this thesis. The edge caching business model can, however, be enhanced by including caches deeper in the network. Research can be extended into the following direction:

- Caches can be placed at internal network nodes by MNOs and their memory space leased to CPs. How would these additional caches affect cache leasing and content placement?

Appendix A

History of Caching

A.1 On Libraries and Encyclopedias

4000 years ago, texts on clay tablet were not unique. Popular texts were continually copied by scribes and subsequently distributed to other collections. This way, they were made available throughout Mesopotamia and ready to be enjoyed and studied by kings and priests without the need to travel to or even conquer neighboring archives.

According to estimates, the Library of Alexandria, founded around 300 BC, contained up to 600,000 scrolls of papyrus. Many works were lent from other parts of the ancient world, and copied and catalogued on arrival. Alexandria is considered revolutionary as a center of science of the ancient world.

The improvement of printing techniques around 1450 allowed the reproduction of texts allowed for the easier reproduction of religious as well as non-religious texts. One of the most ambitious projects using printing techniques was the Encyclopédie (published between 1751 and 1772). Its objective was, among other political goals, to make the world's knowledge available to effectively all the French elite. It is considered that the spread of ideas through

the Encyclopédie contributed to the French Revolution 1789.

These examples show that copy and distributed storage of information reaches back through the ages. Through the placement of copies in various locations, different objectives were achieved: Information was made available for certain publics. Thus, the information was spread more efficiently. Delay of information access as well as danger of information loss were decreased.

A.2 Caches in Computers

In computing and computer networks, hardware or software components that store data in order to serve future requests of that data faster than the main memory are called *caches*. Caches first appeared in research computers in the early 1960s and in production computers later in the same decade; every general-purpose computer built today, from servers to low-power embedded processors, includes caches [PH07, p. 473].

Hardware caches are connected to processing units such as CPUs or GPUs. Here, a cursory overview of the between CPUs and the related caches is given mostly based on Computer Organization and Design by Patterson/Hennessy [PH07, Chapter 7].

The memory hierarchy consists of several levels of memory storage. The higher the memory is in the hierarchy, the lower is its access time (MAT). At the same time, the faster the memory is, the more expensive and thus smaller it is. Caches are located in the memory hierarchy between the CPU registers (highest level) and the random access memory. In modern general-purpose computers, the caches make up three sublevels, L1 to L3. As an example, the minimum access times and sizes of the cache levels in the Intel Skylake microarchitecture (see [Int16]), produced since 2015, are listed in Table A.1.

The data set stored in the higher level memory always is a subset of the lower memory while the lowest level contains all data. When the CPU has

Cache	Size	MAT (cycles)
L1	32 KB	4
L2	256 KB	12
L3	2 MB	44

Table A.1: Intel Skylake cache levels per core [Int16]

a data request, it consecutively searches the memory hierarchy until it finds the data. This way, the illusion of a large and fast memory is created: The amount of data that is frequently requested is relatively small in comparison to the total amount of data in the lowest memory level. This is due to the *temporal* and the *spatial locality* principle. The former describes the probability tha data that was recently requested will soon be requested again. The latter states that if a data location is referenced, data locations with nearby adresses will be referenced soon.

Read request run level by level through the memory hierarchy from top towards bottom until the data is found. From level to level, the *hit time*, i.e. the time it takes to retrieve the requested data, increases. The *miss penalty* for a memory level is the time it takes to retrieve the data from a lower memory level and place the content on all memory levels up to the top. When a miss occurs, the retrieved from the lower level replaces some on the current level. One classic replacement policy is *LRU*, in which the item *least recently used* is replaced.

A.3 Existing Caching Infrastructure

A.3.1 World Wide Web and Network Caches

The World Wide Web is originally designed for end-to-end (E2E) data connections [Hus99]. Content requests are routed from the client's device to the

CP's content storage. From there, the requested content is delivered via the backbone network and the network of the client's Internet Service Provider (ISP) to the client's device. This design has several advantages over a more decentralized approach:

- Coherence: When content is modified at the central storage, the new version will be delivered at every future request.
- Individualized Content: Using some form of security model, clients can be authenticated and receive privileged information.
- Tracking: The IDs and contextual information of the clients can be tracked in a centralized fashion.
- Security: E2E encryption, for example using Transport Layer Security (TLS), can be enabled.

There are, however, also significant drawbacks to the E2E architecture: Single servers as well as data centers providing very popular content are placed under considerable stress. Both the number of simultaneous client connections and the total data throughput are limited. The network surrounding the content storage has traffic limitations as well and can be prone to congestion. Furthermore, the redundant transmission of data to clients in similar locations can create a network overload, and it also incurs financial transmission costs.

With the increasing importance of computer networks in the 1990s, in order to address these drawbacks, the deployment of caching infrastructure became more and more prevalent. www content is cached in various locations of the network. On the one hand, caches are included in browsers making content available locally at the user equipment that has been previously requested. On the other hand, web caches are placed at various positions in the network, between the network edge and the content source. Whenever a client request is passed through a web cache agent, the latter forwards the request to the original source as a proxy for the client. The server's

response is retained in the cache while a copy is transmitted to the client, and, for example, the least recently used (LRU) cached content is replaced. If soon after another request for the same content is passed to the cache agent, the request can be served from the cache instead of its original source. This way, the caches are populated based on previous user requests (pull caching). A different caching paradigm is called *push caching* or *prefetching* [Bes96]. Here, content is placed a priori according to estimated content popularity.

All parties involved benefit from a high cache *hit rate* both in terms of page hit rate (share of requests served from cache) and byte hit rate (page hit rated weighted by file sizes). Assuming that web caches are maintained by an ISP, the following trade-offs are connected to caching:

- **CPs** trade off a reduction of load on data center against less knowledge about its users since they cannot control who downloads cached content.
- **ISPs** save financial cost for file transfer from other network operators and gain better Quality of Service (QoS) for their users but they incur financial cost for installation and maintenance of caches.
- **Client** get improved delay and QoS while some security issues may arise.

A.3.2 Data Centers and Content Delivery Networks

Data Centers

With the rapid growth of Internet traffic in the late 1990s, the combination of E2E delivery and web caching proved to be insufficient. Smaller websites continue to be hosted on central servers that may consist only of one single computer, thus risking service outage in case of physical failure of the server or a request overload. Furthermore, single servers have limitations both regarding the data rate and the number of clients that can be served

simultaneously. Even if server capacity constraints are not a limiting factor, delivery delay may be become an issue in case of greater network distance between client and server.

Large CPs such as Youtube rely on much more involved networks of servers called data centers. Data centers are clusters of servers (or: computational nodes). The purpose of data centers is to handle massive numbers of content requests. Request load can be balanced between different servers. Content access from every server is provided by low latency, high capacity interconnections between the servers. Various architectures for these links are applied in practice [AFLV08].

Content Delivery Networks

Several companies have emerged that combine large scale data centers around the world to Content Delivery Networks (CDNs). CDNs consist of several points of presence (PoPs) at which network caches are deployed, together with switches that provide server-load balancing and request routing. Following [HB05], a CDN is defined as a communication network that deploys infrastructure components operating at protocol Layers 4–7. These components interconnect with each other, creating a virtual network layered on top of an existing packet network infrastructure. Its functions can include

- Content distribution: Services for transmitting the requested content from the original data center or one of the caches to the user.
- Request-routing: Services for navigating user requests to a location best suited for retrieving the requested content.
- Content processing: Adapting the requested content to suit user preferences and device capabilities, i.e. adaptation of content to wireless devices.
- Authorization, authentication, and accounting: Providing user tracking to the CP as well as ensuring user identity to all parties of the file

transmission.

CDNs resolve some of the drawbacks of traditional web caching, in particular regarding individualized content. The security/encryption question can only be resolved if user and CP trust the CDN provider to be an intermediary in their secure communication.

As an example, Google/Youtube owns 15 data centers that are distributed around the world, thus decreasing geographic effects on the delivery delay to clients [Dat18]. In 2017, it was estimated that the total amount of servers in these Google data centers was around 2.5 million. According to Google itself, their recently opened Data Center in Eemshaven (Netherlands) alone was an investment of 600 million € that also includes fast and high capacity links to the surrounding network [Goo18].

Other known CDN providers include Akamai and Amazon Web Services (AWS) that sell CDN services to various content providers. Some CDNs provide the named services only for the content of specific CPs, such as Google Global Cache (service for Youtube) and Facebook Photo CDN (see [PIT+18]).

There are large CPs that rely completely on third-party infrastructure for content delivery. For example, Netflix does not operate any data centers any more but has migrated completely to AWS [Izr16]. Additionally, Netflix itself offers caching infrastructure to network operators called Netflix Open Connect. This project consists of Netflix placing caches either within internet exchange points or inside the networks of ISPs[Net18]. The hardware for these Open Connect Appliances (OCAs) are provided by Netflix while the ISPs provide power, space and connectivity. The cache management remains with Netflix. The company executes cache updates in off-peak hours to bring their content closer to the user and decongest the ISP network.

A.3.3 Cloud and Fog Computing

Services similar to CDNs that offer computational capabilities in addition to data storage are called Cloud Computing. Several large Cloud Computing providers are on the market such as Amazon Elastic Compute Cloud, IBM SmartCloud and Google Compute Engine. The term Fog Computing is used when computational capabilities are distributed and brought closer to the user, e.g. in the context of Internet of Things (IoT). Many results of this work can be transferred from edge caching to edge/fog computing since many requirements overlap.

Appendix B

Peer-reviewed Publications during the Thesis

The work of the following peer-reviewed research papers are incorporated in this thesis:

1. J. Krolikowski, A. Giovanidis, and M. Di Renzo. Fair Distributed User-traffic Association in Cache Equipped Cellular Networks. In WiOpt-CCDWN, pages 1–6, 2017.
2. J. Krolikowski, A. Giovanidis, and M. Di Renzo. Optimal Cache Leasing from a Mobile Network Operator to a Content Provider. In IEEE INFOCOM 2018 – IEEE Conference on Computer Communications, Honolulu, USA, April 2018.
3. J. Krolikowski, A. Giovanidis, and M. Di Renzo. A decomposition framework for optimal edge-cache leasing. IEEE Journal on Selected Areas in Communications, 2018.

The contents in these publications refer to the following aspects of this thesis.

1. The workshop paper [KGR17] investigates load balancing through cache-aware user association.

- The system model and problem statement in Section 4.2 are based on this work. As a savings function, the work considers only (4.6).
- The solution developped in Section 4.4 is a generalization of the algorithm in this publication.
- The numerical evaluation for load-balancing (Section 4.5) is derived from this paper.

2. The conference paper [KGD18] establishes the cache leasing business model that is described here in Chapter 3. Furthermore:
   - The general formulation for the weighted savings function (4.3) in Section 4.2 originates in this paper.
   - The respective system model and problem statement from the CP perspective appear here in Section 5.2.
   - The solution in Section 5.4 is largely based on this work.
   - The numerical evaluation (Section 5.2, in particular 5.6.3 to 5.6.6) is partly based on this article as well.
3. The journal paper [KGDR18] is an extension of [KGD18]. Its original content is:
   - the generalized solution of user association (Section 4.4),
   - the complexity analysis of the cache leasing and content placement (CLCP) problem in Section 5.3 and Corollary3,
   - extended numerical results in Section 5.2, particularly the results discussed in the subsections on policy comparison (5.6.7), sensitivity analysis (5.6.8) and throughput maximization (5.6.9).

Appendix C

Abbreviations and Notation

C.1 Abbreviations

AP	Access Point
AWS	Amazon Web Services
BBU	Baseband Uint
BS	Base Station
CBS	Cache-equipped Base Station
CDN	Content Delivery Network
Closest	User association to the CBS providing the strongest signal
CLCP	Cache Leasing and Content Placement
CoMP	Coordinated Multi-Point
CP	Content Provider
CPU	Central Processing Unit
C-RAN	Cloud Radio Access Network
D2D	Device-to-device
DC	Data Center
E2E	End-to-end
eNB	E-UTRAN Node B

EU	European Union
GDPR	General Data Protection Regulation
GEO	Geographic Caching
GPU	Graphics Processing Unit
HetNet	Heterogeneous Network
HD	Helper Decision problem (from [SGD+13])
HR	Hit Ratio
ICN	Information Centric Networking
IF	Interface
IP	Internet Protocol
IoT	Internet of Things
ISP	Internet Service Provider
LB	Load-balancing
LFU	Least Frequently Used content replacement policy
LRU	Least Recently Used content replacement policy
LTE	Long Term Evolution
MAT	Minimum Access Time
MBS	Macro Base Station
MNO	Mobile Network Operator
MPC	Most Popular Content
Opt- h	User association optimizing a performance criterion represented by savings function h
NUM	Network Utility Maximization
OCA	Open Connect Appliances
PPP	Poisson Point Process
QoS	Quality of Service
RAT	Radio Access Technology
RRH	Remote Radio Head
SBS	Small Base Station

SINR	Signal-To-Interference-and-Noise-Ratio
SW	Social Welfare
TCP/IP	Transmission Control Protocol/Internet Protocol
TLS	Transport Layer Security
TP	Throughput
TTL	Time-to-live
UA	User Association
UE	User Equipment
WiFi	Wireless technology based on the IEEE 802.11x standard
www	World Wide Web

Table C.1: Abbreviations

C.2 Notation

$\mathcal{M}$	Set of CBSs
m	Single CBS
$\mathcal{F}$	Content catalog of typical CP
f	Content file
Π	User association policy, e.g. Opt- h or Closest
$\mathcal{S}^{\Pi}$	Network regions depending on association policy Π
$\mathcal{S}$	Network regions in case association policy is clear
s	Single network region
$x_{m,f}$	decision variable/parameter indicating if file f is placed in CBS m
$\mathbf{x}$	vector of placement variables
$N_{s,f}$	Expected number of CP users in region s requesting file f
$\mathbf{N}$	Content popularity vector
$y_{m,s,f}$	Variable indicating number of users in region s requesting file f associated to CBS m
$\mathbf{y}$	User association vector
$\mathcal{Y}^{\Pi}$	Domain of feasible user association vectors under association policy Π
$\mathrm{h}(\cdot)$	Savings function of the MNO or CP, depending on context: $\mathrm{h}^{\mathrm{MNO}}(\cdot)$, $\mathrm{h}^{\mathrm{CP}}(\cdot)$ or $\mathrm{h}^{(r)}(\cdot)$
$w_{m,s,f}$	Weights referring to user association
$\mathrm{U}_m(\cdot)$	Utility function for CBS m
$\mathbf{w}$	Weight vector referring to user association
$v_m^{\mathbf{w}}(\mathbf{y})$	Weighted traffic volume associated to CBS m with user association $\mathbf{y}$
q	$q = (s, f)$, tuple of region and file, called region-file
$\mathcal{Q}$	Set of region-files

k_m	Number of memory units installed at CBS m
z_m	Variable indicating the number of memory units leased at m
$\mathbf{z}$	Vector of cache leasing variables
$\mathcal{X}$	Domain of feasible tuples $(\mathbf{x}, \mathbf{z})$
b_f	Size of file f
b_{MU}	Size of memory units
p_m	Price per memory unit at CBS m
$\mathcal{R}$	Set of CPs
r	CP, in case several are considered
d_r	Bid by CP r
$\mathbf{d}_{-r}$	Vector of all bids except CP r
$P_p^{(r)}(\cdot)$	Payoff function for price taking CP r for price p
$Q^{(r)}(\cdot)$	Payoff function for price anticipating CP r for price p

Table C.2: Notation

List of Figures

List of Tables

Bibliography

[ACG+13] G. Acs, M. Conti, P. Gasti, C. Ghali, and G. Tsudik. Cache privacy in named-data networking. In *2013 IEEE 33rd International Conference on Distributed Computing Systems*, pages 41–51, July 2013.

[ADR16] A. Araldo, G. Dan, and D. Rossi. Stochastic dynamic cache partitioning for encrypted content delivery. In *Internet Teletraffic Congress (ITC) 2016*, 2016.

[AFLV08] M. Al-Fares, A. Loukissas, and A. Vahdat. A scalable, commodity data center network architecture. In *ACM SIGCOMM Computer Communication Review*, volume 38, pages 63–74. ACM, 2008.

[AGL16] G. Alfano, M. Garetto, and E. Leonardi. Content-centric wireless networks with limited buffers: When mobility hurts. *IEEE/ACM Transactions on Networking*, 24(1):299–311, Feb 2016.

[AGS17] K. Avrachenkov, J. Goseling, and B. Serbetci. A low-complexity approach to distributed cooperative caching with geographic constraints. *Proc. ACM Meas. Anal. Comput. Syst.*, 1(1):27:1–27:25, June 2017.

[And13] J. G. Andrews. Seven ways that hetnets are a cellular paradigm shift. *IEEE Communications Magazine*, 51(3):136–144, March 2013.

[BBD14] E. Baştuğ, M. Bennis, and M. Debbah. Cache-enabled small cell networks: Modeling and tradeoffs. In *Wireless Communications Systems (ISWCS), 2014 11th International Symposium on*, pages 649–653, Aug 2014.

[BBD15] E. Baştuğ, M. Bennis, and M. Debbah. A transfer learning approach for cache-enabled wireless networks. In *2015 13th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, pages 161–166, May 2015.

[Bes96] A. Bestavros. Speculative data dissemination and service to reduce server load, network traffic and service time in distributed information systems. In *Data Engineering, 1996. Proceedings of the Twelfth International Conference on*, pages 180–187. IEEE, 1996.

[BG15] B. Błaszczyszyn and A. Giovanidis. Optimal geographic caching in cellular networks. In *Communications (ICC), 2015 IEEE International Conference on*, pages 3358–3363. IEEE, 2015.

[BGLR93] M. Bellare, S. Goldwasser, C. Lund, and A. Russell. Efficient probabilistically checkable proofs and applications to approximations. In *Proceedings of the Twenty-fifth Annual ACM Symposium on Theory of Computing*, STOC '93, pages 294–304, New York, NY, USA, 1993. ACM.

[BGW10] S. Borst, V. Gupta, and A. Walid. Distributed caching algorithms for content distribution networks. In *Proceedings of*

the 29th Conference on Information Communications, INFOCOM'10, pages 1478–1486, Piscataway, NJ, USA, 2010. IEEE Press.

[BNGP17] B. N. Bharath, K. G. Nagananda, D. Guenduez, and H. V. Poor. Learning-based content caching with time-varying popularity profiles. In *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, pages 1–6, Dec 2017.

[BSW12] A. Brodersen, S. Scellato, and M. Wattenhofer. Youtube around the world: Geographic popularity of videos. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 241–250, New York, NY, USA, 2012. ACM.

[BT89] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1989.

[BW05] D. Bertsimas and R. Weismantel. *Optimization over integers*. Athena Scientific, 2005.

[CBK18] A. Chattopadhyay, B. Błaszczyszyn, and H. P. Keeler. Gibbsian on-line distributed content caching strategy for cellular networks. *IEEE Transactions on Wireless Communications*, 17(2):969–981, Feb 2018.

[CDL+17] W. Chu, M. Dehghan, J. C. S. Lui, D. Towsley, and Z.-L. Zhang. Joint Cache Resource Allocation and Request Routing for In-network Caching Services. *arXiv:1710.11376*, October 2017.

[CHH+17] M. Chen, Y. Hao, L. Hu, K. Huang, and V. K. N. Lau. Green and mobility-aware caching in 5g networks. *IEEE Transactions on Wireless Communications*, 16(12):8347–8361, Dec 2017.

[Chv79] V. Chvatal. A greedy heuristic for the set-covering problem. *Math. Oper. Res.*, 4(3):233–235, August 1979.

[Cis17] Cisco visual networking index: Global mobile data traffic forecast update, 2016–2021 white paper. White Paper, 6 2017.

[CLQK17] Z. Chen, J. Lee, T. Q. S. Quek, and M. Kountouris. Cooperative caching and transmission design in cluster-centric small cell networks. *IEEE Transactions on Wireless Communications*, 16(5):3401–3415, May 2017.

[CTW02] H. Che, Y. Tung, and Z. Wang. Hierarchical web caching systems: Modeling, design and experimental results. *IEEE Journal on Selected Areas in Communications*, 20(7):1305–1314, 2002.

[Dat18] Google Data Center FAQ, 2018.

[DCNT17] M. Dehghan, W. Chu, P. Nain, and D. Towsley. Sharing LRU cache resources among content providers: A utility-based approach. *arXiv:1702.01823*, 2017.

[DEAH17] V. G. Douros, S. E. Elayoubi, E. Altman, and Y. Hayel. Caching games between content providers and internet service providers. *Performance Evaluation*, 2017.

[DJS+17] M. Dehghan, B. Jiang, A. Seetharam, T. He, T. Salonidis, J. Kurose, D. Towsley, and R. Sitaraman. On the complexity of optimal request routing and content caching in heteroge-

neous cache networks. *IEEE/ACM Transactions on Networking*, 25(3):1635–1648, June 2017.

[DMT+16] M. Dehghan, L. Massoulie, D. Towsley, D. Menasche, and YC Tay. A utility optimization approach to network cache design. In *Computer Communications, IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on*, pages 1–9. IEEE, 2016.

[DRGC13] M. Di Renzo, A. Guidotti, and G. E. Corazza. Average rate of downlink heterogeneous cellular networks over generalized fading channels: A stochastic geometry approach. *IEEE Trans. on Comm.*, 61(7):3050–3071, 2013.

[DSJ+15] M. Dehghan, A. Seetharam, Bo Jiang, Ting He, Th. Salonidis, J. Kurose, D. Towsley, and R. Sitaraman. On the complexity of optimal routing and content caching in heterogeneous networks. In *IEEE INFOCOM*, 2015.

[ER15] S. Elayoubi and J. Roberts. Performance and cost effectiveness of caching in mobile access networks. In *Proceedings of the 2nd ACM Conference on Information-Centric Networking*, ACM-ICN '15, pages 79–88, New York, NY, USA, 2015. ACM.

[ETS17] European Telecommunications Standards Institute (ETSI). *ETSI TR 138 913 V14.2.0 (2017-05)*, 10 2017.

[Eur16] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union*, L119:1–88, May 2016.

[Fei98] U. Feige. A threshold of ln n for approximating set cover. *J. ACM*, 45(4):634–652, July 1998.

[FNNT12] N. C. Fofack, P. Nain, G. Neglia, and D. Towsley. Analysis of ttl-based cache networks. In *6th International ICST Conference on Performance Evaluation Methodologies and Tools*, pages 1–10, Oct 2012.

[GA16] A. Giovanidis and A. Avranas. Spatial multi-LRU caching for wireless networks with coverage overlaps. *SIGMETRICS Perform. Eval. Rev.*, 44(1):403–405, June 2016.

[Geo72] A. M. Geoffrion. Generalized Benders decomposition. *Journal of Optimization Theory and Applications*, 10:237–260, 1972.

[GH17] N. Gast and B. Van Houdt. TTL approximations of the cache replacement algorithms LRU(m) and h-LRU. *Performance Evaluation*, 117:33–57, 2017.

[GJS+12] R. A. Gorcitz, Y. Jarma, P. Spathis, M. Dias de Amorim, R.i Wakikawa, J. Whitbeck, V. Conan, and S. Fdida. Vehicular carriers for big data transfers (poster). In *Vehicular Networking Conference (VNC), 2012 IEEE*, pages 109–114. IEEE, 2012.

[GKB12] A. Giovanidis, J. Krolikowski, and S. Brueck. A 0–1 program to form minimum cost clusters in the downlink of cooperating base stations. In *Wireless Communications and Networking Conference (WCNC), 2012 IEEE*, pages 940–945. IEEE, 2012.

[GMDC13] N. Golrezaei, AF Molisch, AlG Dimakis, and G. Caire. Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution. *IEEE Communications Magazine*, 51(4):142–149, 2013.

[Goo18] Google Data Centers – Eemshaven, Netherlands, 2018.

[GXF+14] Y. Guan, Y. Xiao, H. Feng, C. C. Shen, and L. J. Cimini. Mobicacher: Mobility-aware content caching in small-cell networks. In *2014 IEEE Global Communications Conference*, pages 4537–4542, Dec 2014.

[HB05] M. Hofmann and L. R. Beaumont. *Content networking: architecture, protocols, and practice*. Elsevier, 2005.

[Hus99] G. Huston. Web caching. *The Internet Protocol Journal*, 2(3):2–20, 9 1999.

[Int16] Intel. *Intel 64 and IA-32 Architectures Optimization Reference Manual*, June 2016.

[Izr16] Y. Izrailevsky. Completing the Netflix Cloud Migration, 2016.

[JJ18] K. Jaffrès-Runser and G. Jakllari. Pcach: The case for precaching your mobile data. *CoRR*, abs/1807.10051, 2018.

[Kel97] F. Kelly. Charging and rate control for elastic traffic. *Transactions on Emerging Telecommunications Technologies*, 8(1):33–37, 1997.

[KGD18] J. Krolikowski, A. Giovanidis, and M. Di Renzo. Optimal cache leasing from a mobile network operator to a content provider. In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, Honolulu, USA, April 2018.

[KGDR18] J. Krolikowski, A. Giovanidis, and M. Di Renzo. A decomposition framework for optimal edge-cache leasing. *IEEE Journal on Selected Areas in Communications*, 2018.

[KGR17] J. Krolikowski, A. Giovanidis, and M. Di Renzo. Fair distributed user-traffic association in cache equipped cellular networks. In *WiOpt-CCDWN*, pages 1–6, 2017.

[KWW13] J. Krämer, L. Wiewiorra, and C. Weinhardt. Net neutrality: A progress report. *Telecommunications Policy*, 37(9):794–813, 2013.

[LBZL17] J. Liu, B. Bai, J. Zhang, and K. B. Letaief. Cache placement in fog-rans: From centralized to distributed algorithms. *IEEE Transactions on Wireless Communications*, 16(11):7039–7051, Nov 2017.

[LGP+16] Eun-Kyu Lee, Mario Gerla, Giovanni Pau, Uichin Lee, and Jae-Han Lim. Internet of vehicles: From intelligent grid to autonomous cars and vehicular fogs. *International Journal of Distributed Sensor Networks*, 12(9):1550147716665500, 2016.

[LLR+12] T. Lauinger, N. Laoutaris, P. Rodriguez, Th. Strufe, E. Biersack, and E. Kirda. Privacy implications of ubiquitous caching in named data networking architectures. *Technical Report TR-iSecLab-0812-001, ISecLab, Tech. Rep.*, 2012.

[LLS+17] T. Liu, J. Li, F. Shu, M. Tao, W. Chen, and Z. Han. Design of contract-based trading mechanism for a small-cell caching system. *IEEE Transactions on Wireless Communications*, 16(10):6602–6617, Oct 2017.

[LPG+16] M. Leconte, G. Paschos, L. Gkatzikis, M. Draief, S. Vassilaras, and S. Chouvardas. Placing dynamic content in caches with small population. In *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, pages 1–9, April 2016.

[LPQS17] J. Leguay, G. S. Paschos, E. A. Quaglia, and B. Smyth. Cryptocache: Network caching with confidentiality. In *2017 IEEE International Conference on Communications (ICC)*, pages 1–6, May 2017.

[LSLS18] J. Li, S. Shakkottai, J. C. S. Lui, and V. Subramanian. Accurate learning or fast mixing? dynamic adaptability of caching algorithms. *IEEE Journal on Selected Areas in Communications*, pages 1–1, 2018.

[LSLT16] K.-W. Lim, S. Secci, L.Tabourier, and B. Tebbani. Characterizing and predicting mobile application usage. *Computer Communications*, 95:82–94, 2016. Mobile Traffic Analytics.

[LWHS15] R. Lan, W. Wang, A. Huang, and H. Shan. Device-to-device offloading with proactive caching in mobile cellular networks. In *2015 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6, Dec 2015.

[LXL+14] Z. Li, G. Xie, J. Lin, Y. Jin, M. A. Kaafar, and K. Salamatian. On the geographic patterns of a large-scale mobile video-on-demand system. In *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*, pages 397–405, April 2014.

[MAN14] M. A. Maddah-Ali and U. Niesen. Fundamental limits of caching. *IEEE Transactions on Information Theory*, 60(5):2856–2867, 2014.

[MCC+16] A. Mahmood, C. Casetti, C. Chiasserini, P. Giaccone, and J Harri. Mobility-aware edge caching for connected cars. In *Wireless On-demand Network Systems and Services (WONS), 2016 12th Annual Conference on*, pages 1–8. IEEE, 2016.

[MMPC17] M. Mangili, F. Martignon, S. Paris, and A. Capone. Bandwidth and cache leasing in wireless information-centric networks: A game-theoretic study. *IEEE Transactions on Vehicular Technology*, 66(1):679–695, Jan 2017.

[MSR+17] A. Mondal, S. Sengupta, BR Reddy, MJV Koundinya, Ch. Govindarajan, P. De, N. Ganguly, and S. Chakraborty. Candid with youtube: Adaptive streaming behavior and implications on data consumption. In *NOSSDAV'17*, Taipei, Taiwan, June 2017.

[MW00] J. Mo and J. Walrand. Fair end-to-end window-based congestion control. *IEEE/ACM Trans. Netw.*, 8(5):556–567, October 2000.

[MWZ+17] G. Ma, Z. Wang, M. Zhang, J. Ye, M. Chen, and W. Zhu. Understanding performance of edge content caching for mobile video streaming. *IEEE Journal on Selected Areas in Communications*, 35(5):1076–1089, May 2017.

[MZS+13] A. Mohaisen, X. Zhang, M. Schuchard, H. Xie, and Y. Kim. Protecting access privacy of cached contents in information centric networks. In *Proceedings of the 8th ACM SIGSAC Symposium on Information, Computer and Communications Security*, ASIA CCS '13, pages 173–178, New York, NY, USA, 2013. ACM.

[NCM17] G. Neglia, D. Carra, and P. Michiardi. Cache policies for linear utility maximization. In *2017 IEEE Conference on Computer Communications, INFOCOM 2017, Atlanta, GA, USA, May 1-4, 2017*, pages 1–9, 2017.

[NEHA08] BL Ng, JS Evans, SV Hanly, and D. Aktas. Distributed downlink beamforming with cooperative base stations. *IEEE Transactions on Information Theory*, 54(12):5491–5499, 2008.

[Net18] Netflix open connect overview, 2018.

[New05] MEJ Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351, 2005.

[NLPW14] H. Niu, C. Li, A. Papathanassiou, and G. Wu. Ran architecture options and performance for 5g network evolution. In *2014 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, pages 294–298, April 2014.

[NMB+15] K.P. Naveen, L. Massoulie, E. Baccelli, A. Carneiro Viana, and D. Towsley. On the interaction between content caching and request assignment in cellular cache networks. In *ACM*, AllThingsCellular '15, pages 37–42. ACM, 2015.

[NRTV07] Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V Vazirani. *Algorithmic game theory*, volume 1. Cambridge University Press Cambridge, 2007.

[PAG13] H. Pinto, J. M Almeida, and M. A. Gonçalves. Using early view patterns to predict the popularity of youtube videos. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 365–374. ACM, 2013.

[Par15] Parliament and Council of European Union. Regulation (EU) no 2015/2021. `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2015.310.01.0001.01.ENG`, 2015. [Online; accessed 19-July-2018].

[PBL+16] G. Paschos, E. Başţuğ, I. Land, G. Caire, and M. Debbah. Wireless caching: Technical misconceptions and business barriers. *IEEE Communications Magazine*, 54(8):16–22, 2016.

[PH07] D. A. Patterson and J. L. Hennessy. *Computer Organization and Design: The Hardware/Software Interface*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2007.

[PIP+16] K. Poularakis, G. Iosifidis, I. Pefkianakis, L. Tassiulas, and M. May. Mobile data offloading through caching in residential 802.11 wireless networks. *IEEE Transactions on Network and Service Management*, 13(1):71–84, 2016.

[PIST16] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas. Exploiting caching and multicast for 5g wireless networks. *IEEE Transactions on Wireless Communications*, 15(4):2995–3007, April 2016.

[PIT14] K. Poularakis, G. Iosifidis, and L. Tassiulas. Approximation algorithms for mobile data caching in small cell networks. *IEEE Transactions on Communications*, 62(10):3665–3677, 2014.

[PIT+18] G. S. Paschos, G. Iosifidis, M. Tao, D. Towsley, and G. Caire. The Role of Caching in Future Communication Systems and Networks. *ArXiv e-prints*, May 2018.

[PMGEA16] F. De Pellegrini, A. Massaro, L. Goratti, and R. El-Azouzi. A pricing scheme for content caching in 5g mobile edge clouds. In *2016 International Conference on Wireless Networks and Mobile Communications (WINCOM)*, pages 193–198, Oct 2016.

[PT13] K. Poularakis and L. Tassiulas. Exploiting user mobility for wireless content delivery. *2013 IEEE International Symposium on Information Theory*, pages 1017–1021, 2013.

[PT17] K. Poularakis and L. Tassiulas. Code, cache and deliver on the move: A novel caching paradigm in hyper-dense small-cell networks. *IEEE Transactions on Mobile Computing*, 16(3):675–687, March 2017.

[Rus95] A. Ruszczyński. On convergence of an augmented lagrangian decomposition method for sparse convex optimization. *Mathematics of Operations Research*, 20(3):634–656, 1995.

[Sch86] A. Schrijver. *Theory of Linear and Integer Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1986.

[SGD+13] K. Shanmugam, N. Golrezaei, A.G. Dimakis, A.F. Molisch, and G. Caire. Femtocaching: Wireless content delivery through distributed caching helpers. *Information Theory, IEEE Transactions on*, 59(12):8402–8413, Dec 2013.

[ST85] DD Sleator and RE Tarjan. Amortized efficiency of list update and paging rules. *Communications of the ACM*, 28(2):202–208, 1985.

[TAG+15] S. Traverso, M. Ahmed, M. Garetto, P. Giaccone, E. Leonardi, and S. Niccolini. Unravelling the impact of temporal and geographical locality in content caching systems. *IEEE Transactions on Multimedia*, 17(10):1839–1854, Oct 2015.

[TNS17] A. Tuholukova, G. Neglia, and T. Spyropoulos. Optimal cache allocation for femto helpers with joint transmission capabilities.

In *2017 IEEE International Conference on Communications (ICC)*, pages 1–7, May 2017.

[WCT+14] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. Leung. Cache in the air: exploiting content caching and delivery techniques for 5G systems. *IEEE Communications Magazine*, 52(2):131–139, 2014.

[WJP+13] S. Woo, E. Jeong, Sh. Park, J. Lee, S. Ihm, and KS Park. Comparison of caching strategies in modern cellular backhaul networks. In *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '13, pages 319–332, New York, NY, USA, 2013. ACM.

[WSH15] T. Wang, L. Song, and Z. Han. Dynamic femtocaching for mobile users. In *2015 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 861–865, March 2015.

[WZSL17] R. Wang, J. Zhang, SH Song, and KB Letaief. Mobility-aware caching in d2d networks. *IEEE Transactions on Wireless Communications*, 16(8):5001–5015, 2017.

[XvdSLL15] J. Xu, M. van der Schaar, J. Liu, and H. Li. Forecasting popularity of videos using social media. *IEEE Journal of Selected Topics in Signal Processing*, 9(2):330–343, March 2015.

[YCXW15] C. Yang, Z. Chen, B. Xia, and J. Wang. When icn meets c-ran for hetnets: an sdn approach. *IEEE Communications Magazine*, 53(11):118–125, November 2015.

[Yoo17] CS Yoo. Wireless Network Neutrality: Technological Challenges and Policy Implications. *Berkeley Technology Law Journal*, 31, 2017.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Gestion de contenu optimale et dimensionnement de memoire dans les réseaux sans fil

Mots clés : Contenu, Réseau sans fil, Cache, Station de base, Mémoire, Planification

Résumé : L'augmentation massive du trafic cellulaire pose de sérieux défis à tous les acteurs concernés par la diffusion de contenu sans fil. Alors que la densification du réseau permet d'accéder à des utilisateurs supplémentaires, les liaisons de transport à grande vitesse et à grande capacité sont coûteuses. La mise en cache du contenu populaire au bord du réseau promet permettre de décharger le trafic utilisateur de ces connexions, susceptibles d'être encombrées, ainsi que des centres de données du réseau fédérateur.
Cette thèse propose un modèle commercial dans lequel un opérateur de réseau mobile (MNO) préinstalle et entretient des caches sur son équipement sans fil (stations de base avec cache, CBS). L'espace mémoire ainsi que les capacités de calcul sont ensuite loués aux fournisseurs de contenu (CP) qui souhaitent rapprocher leur contenu à l'utilisateur. Pour une compensation financière, un CP peut alors décharger le trafic de son centre de données et améliorer la qualité de service des utilisateurs. Le CP prend des décisions de placement de contenu en fonction des données prédictives sur le trafic des utilisateurs et la popularité du contenu. Dans la phase de livraison, les utilisateurs peuvent être desservis à partir des caches au cas où ils seraient associés à des stations sur lesquelles le contenu demandé est mis en cache.
Ce travail examine trois aspects du modèle commercial proposé: La première question de recherche porte sur l'association des utilisateurs en tant qu'élément central du schéma de mise en cache au bord du réseau. Les stratégies d'association des utilisateurs prenant en charge le cache peuvent permettre aux utilisateurs des zones de chevauchement de couverture d'être associés à une CBS contenant le contenu demandé plutôt que conventionnellement à celui qui fournit le signal le plus puissant. La thèse propose un algorithme décentralisé original pour l'association d'utilisateurs appelée Generalized Bucket-filling qui permet des gains au-delà de la maximisation du taux de réussite (hit ratio). Les mesures de performance telles que le débit du réseau et l'équilibrage de la charge des utilisateurs parmi les CBS sont prises en compte. Des expériences montrent que l'association des utilisateurs au cache a) augmente le taux de réussite b) sans surcharger les CBS uniques c) tout en fournissant un débit élevé du système.
Le deuxième problème traité concerne un seul CP qui doit décider combien d'espace de cache à louer à chaque CBS pour un prix fixe et du contenu à placer. Ses choix doivent être basés sur des estimations de la popularité des fichiers ainsi que sur la politique d'association des utilisateurs MNO. Le problème de leasing et de placement du contenu du cache est formulé sous la forme d'un problème non linéaire à nombres entiers mixtes (NLMIP). Dans sa solution, le problème est séparé en un sous-problème CP linéaire discret et un sous-problème continu non linéaire utilisant la décomposition de Benders. Le CP et le MNO coopèrent, aidant le CP à prendre des décisions optimales qui profitent aux deux parties: Le CP maximise ses économies grâce à la mise en cache tandis que le MNO peut trouver le prix de cache optimal et recevoir la compensation financière maximale.
Une troisième question de recherche élargit la portée de l'interaction entre plusieurs CPs et un opérateur de réseau mobile. Désormais, le MNO ne fixe pas de prix fixe par unité de mémoire, mais réagit aux demandes du CP en matière d'espace mémoire en fonction des économies réalisées grâce à la mise en cache.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Title : Optimal Content Management and Dimensioning in Wireless Networks

Keywords : Content, Wireless Network, Cache, Base Station, Memory, Scheduling

Abstract : The massive increase in cellular traffic poses serious challenges to all actors concerned with wireless content delivery. While network densification provides access to additional users, high-speed and high-capacity backhaul connections are expensive. Caching popular content at the network edge promises to offload user traffic from these congestion prone connections as well as from the data centers in the backbone network.
This thesis proposes a business model in which a mobile network operator (MNO) pre-installs and maintains caches at its wireless equipment (Cached Base Stations, CBSs). Memory space together with computational capabilities is then leased to content providers (CPs) that want to bring their content closer to the user. For a financial compensation, a CP can then offload traffic from its data center and improve user Quality of Service. The CP makes content placement decisions based on user traffic and content popularity data. In the delivery phase, users can be served from the caches in case they are associated to stations that have the requested content cached.
This work investigates three aspects of the proposed business model: The first research question focuses on user association as a central element to the edge caching scheme. Cache-aware user association policies can allow for users in coverage overlap areas to be associated to a CBS that holds the requested content rather than conventionally to the one that provides the strongest signal. The thesis proposes an original decentralized algorithm for user association called Generalized Bucket-filling that allows gains beyond maximizing the hit ratio. Performance metrics such as network throughput and load balancing of users among CBSs are taken into account. Experiments show that cache-aware user association a) increases the hit ratio b) without overloading single CBSs while c) providing high system throughput.
The second problem treated considers a single CP that needs to decide how much cache space to lease at each CBS for a fixed price, and what content to place. Its choices should be based on estimates of file popularity as well as MNO user association policy. The cache leasing and content placement problem is formulated as a non-linear mixed-integer problem (NLMIP). In its solution, the problem is separated into a linear discrete CP subproblem and a non-linear continuous subproblem using Benders decomposition. The CP and the MNO cooperate, helping the CP to make optimal decisions that benefit both parties: The CP maximizes its savings from caching while the MNO can find the optimal cache price and receive the maximum financial compensation.
A third research question widens the focus to the interaction between several CPs and one MNO. Now, the MNO does not set a fixed price per memory unit but instead reacts to CP demands for memory space that depend on the savings they can achieve from caching.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France