Beamforming in Wireless Coded Caching Systems

Master's thesis in Communication Engineering

SNEHA MADHUSUDAN

DEPARTMENT OF ELECTRICAL ENGINEERING

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2022
www.chalmers.se

Master's thesis 2022

Beamforming in Wireless Coded Caching Systems

SNEHA MADHUSUDAN

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Electrical Engineering
Wireless Systems Division
Communication Systems Group
Chalmers University of Technology
Gothenburg, Sweden 2022

Beamforming in Wireless Coded Caching Systems

SNEHA MADHUSUDAN

Supervisors: Behrooz Makki, Ericsson Research, Tommy Svensson, Charitha Madapatha, Hao Guo, Department of Electrical Engineering, Chalmers
Examiner: Tommy Svensson, Department of Electrical Engineering, Chalmers

Master's Thesis 2022
Department of Electrical Engineering
Communication Systems Group
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 767411049

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2022

Beamforming in Wireless Coded Caching Systems
SNEHA MADHUSUDAN
Department of Electrical Engineering
Chalmers University of Technology

Abstract

A design of beamforming along with coded-caching in wireless networks is considered, where a server with multiple antennas broadcasts contents to cache nodes serving the users. This design benefits from the gains obtained from the coded-caching scheme and beamforming while transferring contents to the cache nodes. A comparison of the gains in a network having traditional/un-coded caching scheme is studied. The joint design of beamforming and coded caching provides multi-casting opportunities, which balances out the effect of noise and interference in the wireless network. Finally, the effect of different buffering and decoding methods on the performance of coded-caching scheme is also studied. The results show that with proper beamforming, the coded-caching scheme can reduce the peak backhaul traffic considerably.

Keywords: Multicast, beamforming, coded-caching, un-coded caching, caching, backhaul, 6G, beyond 5G, genetic algorithm and backhaul traffic.

Acknowledgements

Firstly, I would like to thank professor Tommy Svensson for giving me this opportunity to perform the thesis along with him and his team. I also thank Dr. Behrooz Makki for giving me a chance to be a part of the Ericsson family where, the atmosphere and the people that I encountered were highly encouraging and motivated me to be a better researcher and an engineer.
I have learnt a lot in the past few months and I would like to offer my special gratitude to Hao and Charitha for their patience and constant support.
Finally, I would like to thank my family for letting me follow my dream of getting a Masters' degree and for their constant guidance over these two years.

Sneha Madhusudan, Gothenburg, August 2022

List of Acronyms

ADC	Analog to Digital Converter
AWGN	Additive White Gaussian Noise
BS	Base Station
CC	Coded Caching
DAC	Digital to Analog Converter
DFT	Discrete Fourier Transform
D2D	Device-to-Device
E2E	End-to-End
GA	Genetic Algorithm
HT	High Traffic
IAB	Integrated Access and Backhaul
LT	Low Traffic
mmWave	Millimeter Wave
MAN	Mohamed Ali and Nisen
M2M	Machine-to-Machine
MRC	Maximum Ratio Combining
RF	Radio Frequency
SIC	Successive Interference Cancellation
SINR	Signal-to-interference-plus-Noise-Ratio
SNR	Signal-to-Noise-Ratio
STP	Successful Transmission Probability
UC	Uncoded Caching
V2X	Vehicle-to-Everything

Nomenclature

Below is the nomenclature of indices, sets, parameters, and variables that have been used throughout this thesis.

Indices

i	Indices for users
t	Index for time step

Parameters

P	Maximum transmit power
γ	Probability of successfully removing interference
η	Probability of successfully decoding a sub-packet
α	Power split parameter
Z	AWGN

Variables

$\boldsymbol{h_i^{\mathrm{LT}}}$	Channel realizations during low traffic period
$\boldsymbol{h_i^{\mathrm{HT}}}$	Channel realizations during high traffic period
$\boldsymbol{w_i}$	Beamforming weights
N	Number of files
R_i^N	Rate for particular sub-file
R^N	Total Rate for a particular file
K^N	Total information in nats for a particular file
K_i^N	Total information in nats for particular a sub-file
N_i	Sub-files
C_i	Cache node
M	Number of files in a cache node
K	Number of cache nodes

Contents

List of Figures

List of Tables

1
Introduction

1.1 Introduction

According to [1], the concentration of users and the number of wireless devices have increased by orders of magnitude over the recent years. The continuous increase in the traffic within the networks and the high demands for the required data rates by the users [2] are being satisfied by incorporating techniques like network densification as mentioned in [3], [4], [5]. Network densification simply means adding more base stations (BSs) of different types in a specific area to help serve more devices in the network [6]. The main challenge in this technique arises when the number of BSs and the users increases and all BSs have to be connected to the core network through the transport networks [7]. In particular, the increase in the traffic between the BS and a core network, popularly known as the backhaul traffic, may lead to backhaul congestion which in turn causes an end-to-end (E2E) delay in the network. In order to help reduce the backhaul load and E2E latency, wireless caching schemes may be used [8].

Caching is a technique used in wireless systems to store popular, reusable contents nearer to the end users during the low traffic (LT) periods, to help reduce the backhaul load. This technique is found very helpful in backhaul or delay constrained applications like device-to-device (D2D) [9]-[10], vehicle-to-everything (V2X) [11] and integrated access and backhaul (IAB) [12, 13, 14, 15]. In particular, the technique of coded-caching introduced by Maddah-Ali and U. Niesen in [16, 17, 18] has shown promising results in reducing the requirement of the peak rate of the backhaul links during the high traffic (HT) periods.

Along with the concept of coded caching (CC), a technique known as beamforming is also utilized in the thesis to investigate on how beamforming can affect the performance of the CC networks. Beamforming is a signal processing technique that helps improve the signal quality and the user's achievable rate [19].

1.2 Objectives

The aim of this thesis is as follows.

- Perform a deep review and a comparison between un-coded and coded caching schemes.
- Develop a wireless channel model by taking into consideration different wireless aspects of the link between the server and the cache nodes.
- Perform simulations to demonstrate the efficiency of CC, as compared to un-coded caching scheme, and analyse the effect of beamforming.
- Study the effect of various parameters and different message decoding/buffering methods on the network performance.

2

Theory

2.1 Background

According to [20], CISCO predicted that by 2022, 79% of the world's mobile traffic in the network would be video as shown in Fig. 2.1. Therefore, while most of the network traffic like breaking news, trending tweets and video are cacheable, the main importance is given to video.

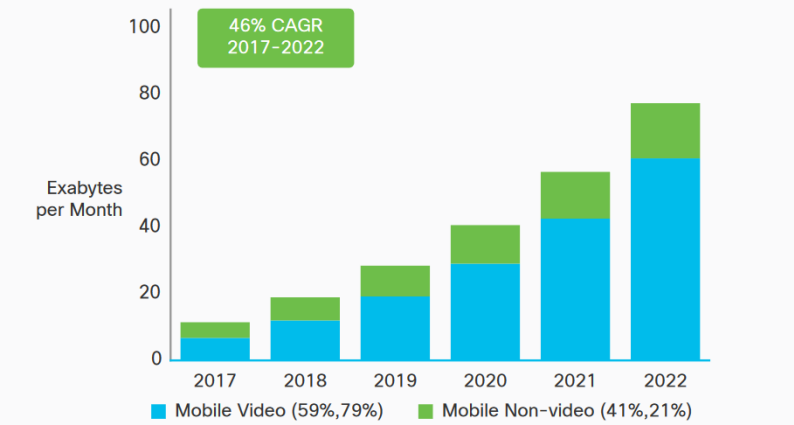

Figure 2.1: *World's mobile data traffic [20].*

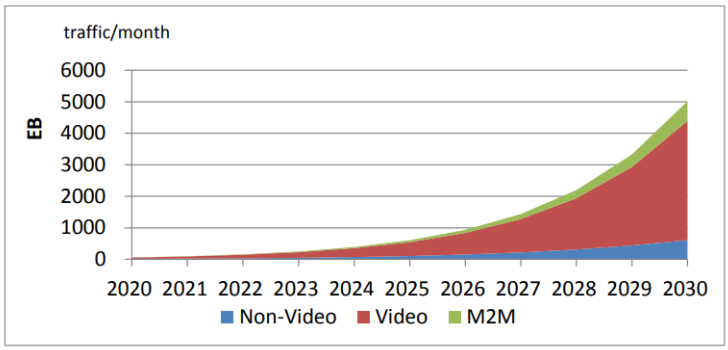

Figure 2.2: *Global mobile traffic by different service types [21].*

Similarly, [21] shows the estimation of global mobile traffic by different service types like video, non-video, machine-to-machine (M2M). It is observed from Fig. 2.2 that the video streaming occupies and contributes to most of the network and is estimated

to increase dramatically after the year 2020. As estimated in [21], the video traffic would be 6 times more than that of the non-video types of service or application in the year 2030, when compared to the year 2020.

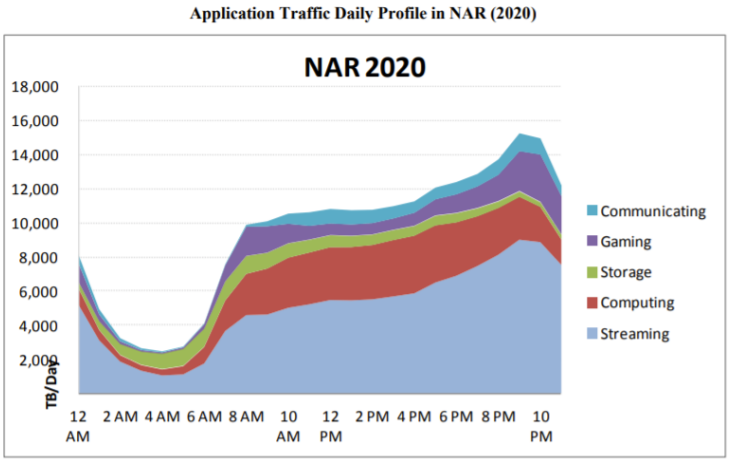

Figure 2.3: *Application traffic daily profile in Western Europe 2020 [21].*

From Fig. 2.3, it is also observed that video has the highest variation in the daily profile and the backhaul peak rate is the highest during the night time, e.g., around 10 pm in Western Europe. The high variations in the daily traffic profile leads to the network not being utilized completely during the off-peak hours, for instance, during the day. This makes the network economically less feasible.

In order to avoid these peak variations and to *flatten out* the curve of Fig. 2.3 or distribute the traffic uniformly during the day, caching schemes are used. The popular contents requested by the users during the high traffic period are predicted beforehand and placed at cache nodes nearer to the end users during the off peak/low traffic hours. The process of caching not only helps in reducing the cost, but also reduces the backhaul load and uniformly distributes the traffic over time to utilize the network in a more efficient manner. Presently, the concept of caching is being used in data delivery systems for example, in Netflix's video streaming service [22] and Facebook's photo caching [23].

2.1.1 Concept of Caching

Caching simply means to predict the popular contents that the users may request during the HT period and store them at the cache nodes nearer to the end users in order to reduce the backhaul congestion and E2E delay in the network. The gains obtained through the process of caching depends highly on the local memory size at the cache nodes and also on the accuracy of predicting the popular contents. For instance, if the size of the local memory of the cache nodes is insufficient to store a good portion of the popular contents, the caching becomes impractical. This leads

to a gain that is almost insignificant [24].

Although the users have memory, the size of the memory in the mobile devices is still not sufficient to store a significant portion of popular contents. Even so, since the number of mobile devices and users are continuously growing, taking advantage of the aggregated cache that is spread across the network would be useful to make it work like a larger cache, which can lead to a significant gain in backhaul peak traffic reduction [24].

The process of caching takes place in two different phases:

- **Placement phase:** This phase of caching occurs during the off peak hours or during the LT period, for instance, around 4 am in the morning. The popular contents, which the users might later request for, are predicted and stored at the cache nodes in this phase.
- **Delivery phase:** This phase occurs during the peak traffic hours or during the HT period, e.g., around 9 pm, when the network is busy. In this phase, the users reveal their requests for the desired contents. If the user's requested content is already cached during the placement phase, the content is delivered to the user directly from the cache node without having to connect to the server. If the requested content is not present in the cache node, it has to be provided from the server.

The event when the user's requested content is not present in the cache node is known as *cache miss*. The difference between the two caching schemes, traditional and coded caching schemes, depends on how the contents are placed at the cache nodes during the placement phase and later accessed during the delivery phase.

2.2 Comparison between Un-Coded Caching and Coded Caching Schemes

This section provides an in-depth conceptual understanding and a comparison between different caching schemes, that is, coded and un-coded caching schemes.

2.2.1 Traditional/Un-Coded Caching

During the placement phase of the traditional/uncoded caching scheme, identical copies/portions of contents are placed at every cache node in the network. In such a scenario, when the users reveal their requests and the requested data is already cached, the data is directly fetched from the local copies made at the cache node during the placement phase, rather than getting them all the way up from the server. Therefore, caching the contents beforehand saves time, makes use of the resources in an efficient manner and reduces the backhaul load. However, when the user requests for a content that is not present in its associated cache node during the delivery phase, the contents have to be fetched all the way up from the server which can cause latencies and results in increased traffic.

The gain obtained in the network following this scheme is dependent on the individual memory size at each cache node. The limitation of this traditional caching scheme is that, if the size of the cache memory at each cache node is insufficient for it to store an ample portion of the popular contents, the obtained gain will be negligible.

2.2.1.1 Un-Coded Caching: Example 1

Consider a simple network having a server with $N = 2$ equally-popular files (File A and File B) and $K = 2$ cache nodes. Let us assume that each user can cache up to $M = 1$ file during the placement phase and each user requests for a single file during the delivery phase. The files in the server are divided into smaller sub-files as A_1, A_2 and B_1, B_2 respectively.
Placement phase: During the placement phase, two cache nodes are filled with identical sub-files of equal size as shown in Fig. 2.4. That is, both the cache nodes are filled with A_1 and B_1.

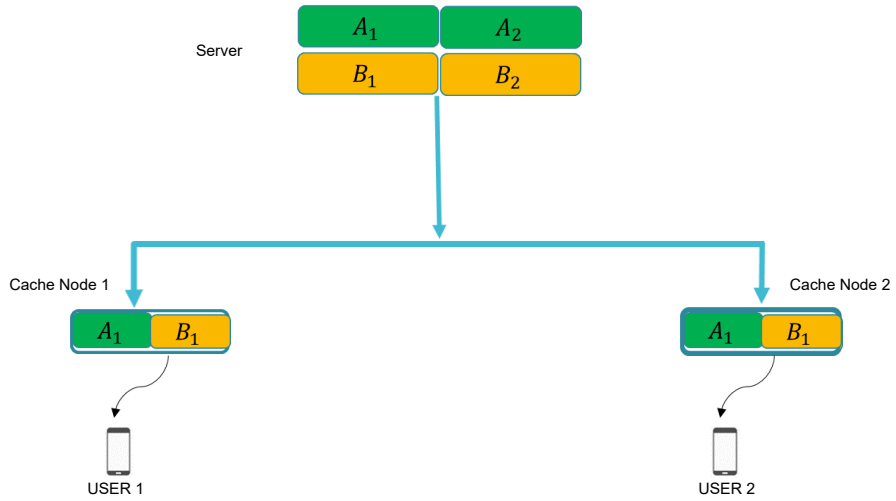

Figure 2.4: *Traditional Caching Scheme: Placement phase.*

Delivery Phase: During the delivery phase, assume User 1 (mobile device 1), associated with cache node 1, requests for file A (green) and User 2 (mobile device 2), associated with cache node 2, requests for file B (yellow) as depicted in Fig. 2.5. The cache nodes are first checked if any portion of requested file is present in the cache. If it is present, the request is satisfied by fetching part of the contents directly from the cache nodes. If the requested file/sub-file is not present in the cache, they are fetched from the server and the cache is possibly updated. The user is then served with its requested contents.

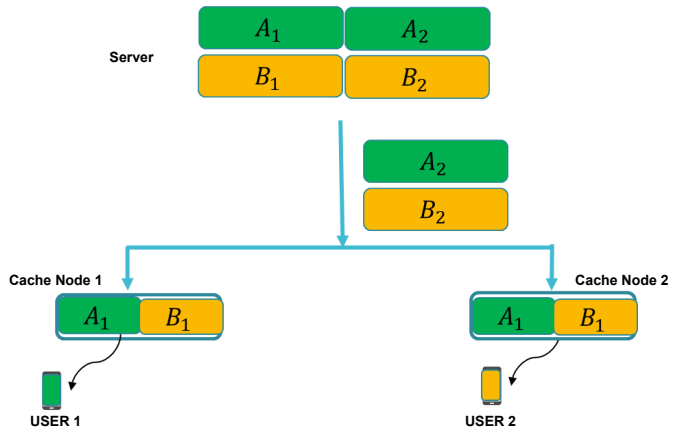

Figure 2.5: *Traditional Caching Scheme: Delivery phase.*

For example, in Fig. 2.5 we can see that User 1 is requesting for file A (green). When checked at the cache node 1, only a part of file A, that is, sub-file A_1, is present in the cache node 1 from the placement phase. Therefore, User 1 receives A_1 directly from the cache node and the complementary sub-file A_2 is fetched from the server (either directly from the sever to the user or via the cache node). User 1 will then have the sub-files A_1 and A_2 which is nothing but file A. Similarly, when User 2 requests for file B (yellow), cache node 2 is checked for any sub-files of B. Since B_1 is already present in the cache node from the placement phase, User 2 receives sub-file B_1 directly from the cache node. The complementary sub-file B_2 is fetched from the server and sent to User 2 (either directly from the sever to the user or via the cache node) so that, in the end, User 2 will have file B.

This thesis aims to reduce the *worst-case* peak traffic load that occurs when different users request for different files. The worst-case peak traffic load obtained in this example using the traditional caching scheme is 1 file. This is because, in the worst-case, one whole file (sub-files A_2 and B_2) is fetched from the server during the delivery phase. Suppose both the users request for the same file, that is, if both User 1 and User 2 request for file A (yellow), only sub-file A_2 needs to be broadcast to both the users during the delivery phase. The peak traffic in this scenario would then be half a file.

The worst-case peak traffic load in the network obtained for the considered example of traditional caching scheme is given by:

$$R(M) = K\left(1 - \frac{M}{N}\right) = 2\left(1 - \frac{1}{2}\right) = 1, \tag{2.1}$$

where, $(1 - \frac{M}{N})$ in (2.1) is known as the local caching gain [25] and is dependent on the individual cache sizes in the network. Therefore, if the cache size is not large enough to store a significant portion of contents, the gain obtained becomes insignificant [24].

When compared to the no-cache scheme, that is, a network without any caching, the worst-case peak traffic load is improved by the traditional caching scheme. In

a network without any caching, the worst-case peak traffic load for the above considered example is 2 files. That is, two whole files will need to be fetched all the way from the server in order to satisfy the users with their requested contents or files.

2.2.1.2 Un-Coded Caching: Example 2

Consider another example of a network with $N = 3$ equally-popular files and $K = 3$ cache nodes. Assume that each user can cache up to $M = 1$ file during the placement phase and each user requests for a single file during the delivery phase. The 3 files are further divided into three sub-files that are concatenated as:

$$\begin{aligned} A &- - > A_1, A_2, A_3, \\ B &- - > B_1, B_2, B_3, \\ C &- - > C_1, C_2, C_3. \end{aligned} \tag{2.2}$$

Placement phase: During the placement phase, the three cache nodes are filled with identical sub-files of equal size as shown in Fig. 2.6. That is, all cache nodes are filled with sub-files (A_1,B_1,C_1).

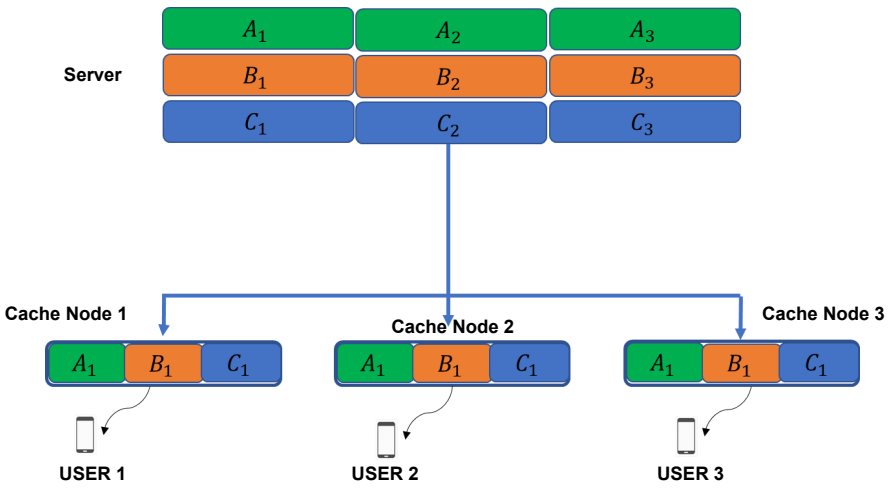

Figure 2.6: *Traditional Caching Scheme example: Placement phase.*

Delivery Phase: During the delivery phase, assume User 1 (mobile device 1), associated with cache node 1, requests for file A (green), User 2 (mobile device 2), associated with cache node 2, requests for file B (orange) and User 3 (mobile device 3), associated with cache node 3 requests for file C (blue) as depicted in Fig. 2.7. As mentioned earlier, the cache nodes are first checked if the requested file is present in the cache. If it is present, the request is satisfied by fetching the contents directly from the cache nodes. If the requested file/sub-file is not present in the cache, it is fetched from the server and then sent to the user.

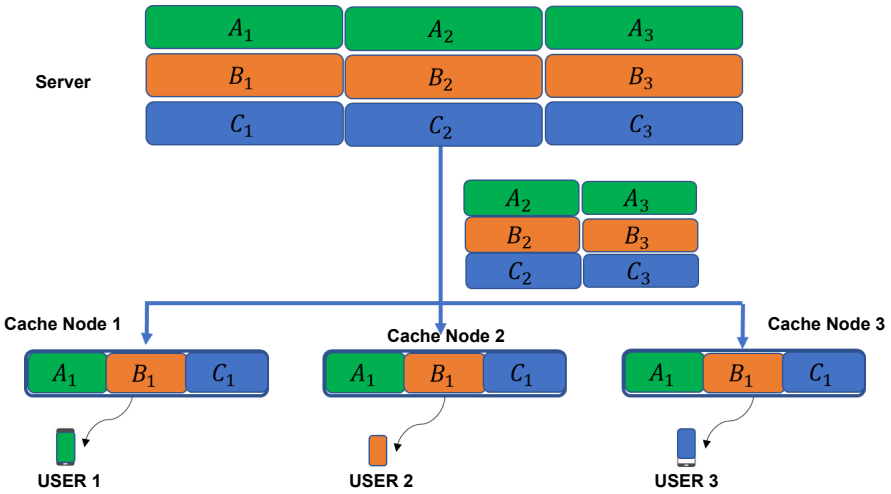

Figure 2.7: *Traditional Caching Scheme example: Delivery phase.*

For example, in Fig. 2.7 it can be seen that User 1 is requesting for file A (green). When checked at the cache node, only a part of file A, that is, sub-file A_1, is present in the cache node from the placement phase. Therefore, the other sub-files A_2 and A_3 are fetched from the server and sent to User 1 during the delivery phase (either directly from the sever to the user or via the cache node). Sub-file A_1 is fetched directly from the cache node as it was cached during the placement phase. The User 1 will then have the sub-files A_1, A_2 and A_3 which is nothing but file A. Similarly, when User 2 requests for file B (orange), the cache node is checked for any file/sub-files of B. Since B_1 is already present in the cache node from the placement phase, B_2 and B_3 are fetched from the server and then all sub-files are concatenated and sent to User 2 during the delivery phase. Finally, when User 3 requests for file C (blue), cache node of User 3 is checked for any sub-file of C. Since C_1 is already present in the cache node from the placement phase, C_2 and C_3 are fetched from the server and then all sub-files are concatenated and sent to User 3 during the delivery phase so that the User 3 is provided with the complete file C.

In this example, the worst-case peak load in the network, i.e., for the users requesting different files, is 2 files. This is because as shown in Fig. 2.7, two whole files, i.e., 6 sub-files each of size 1/3 of a file, are fetched from the server during the delivery phase/high traffic period. This can be calculated as:

$$R(M) = K\left(1 - \frac{M}{N}\right) = 3\left(1 - \frac{1}{3}\right) = 2. \tag{2.3}$$

When the number of users in the network increases, the worst-case peak load also increases. The worst-case peak traffic load in the network with un-coded caching scheme according to [16], [17], [18] is given by

$$R(M) = K\left(1 - \frac{M}{N}\right). \tag{2.4}$$

2.2.2 Coded Caching

The CC scheme focuses on solving the question of how to make use of all the memories in the network and access the contents stored at a cache node during the placement phase, when requested by the users during the delivery phase so that the peak traffic in the network is reduced.

During the placement phase of the CC scheme, different parts of the content files are distributed in different cache nodes, all over the network. Maddah-Ali and Niesen in [16], [17], [26] showed that making use of these distributed contents at different caches in a cooperative manner can result in substantial global gain that corresponds to the total cache available in the network. During the delivery phase, the distributed contents also make way for multi-casting opportunities. That is, a single coded packet that is dependent on the multiple requests from multiple users can be sent in a single transmission which then serves multiple users at once. The most important feature of any CC scheme is to combine all requested packets and form a single packet from which the receivers on the caches can mitigate the interfering packets and retrieve or decode the packet of interest with the help of the already cached sub-packets.

2.2.2.1 Coded Caching: Example 1

Consider again a simple network system having a server with $N = 2$ equally-popular files (File A and File B) and $K = 2$ cache nodes. Let us assume that each user can cache up to $M = 1$ file during the placement phase and each user requests for a single file during the delivery phase. The files in the server are divided into smaller sub-files as A_1, A_2 and B_1, B_2.

Placement phase: During the placement phase, different contents are placed at different cache nodes all over the network and the network is diversified as shown in Fig. 2.8. The sub-files A_1, B_1 are placed at cache node 1 and sub-files A_2, B_2 are placed at cache node 2. For the sake of simplicity, we assume that all sub-files to be of equal size. When compared to the un-coded caching scheme, the transmission load of sub-files during the placement phase of the CC scheme is doubled (see Section 2.2.1.1). However, this is not problematic as the process of placement occurs during the LT period.

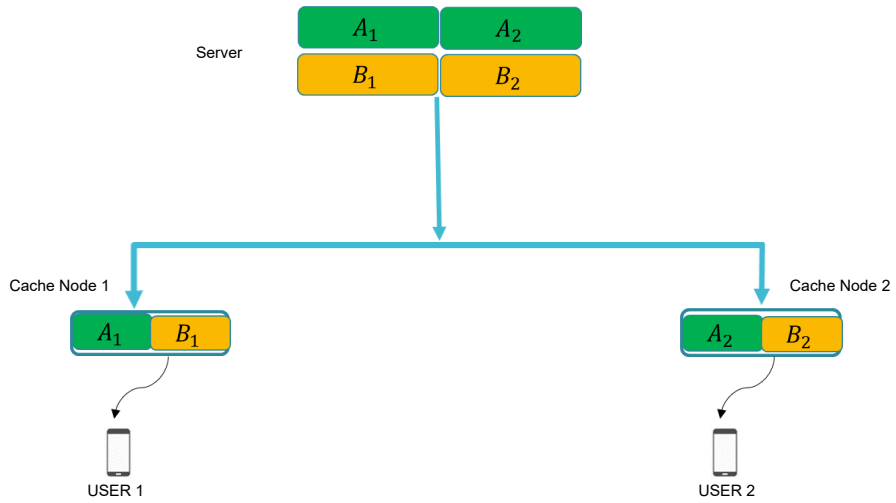

Figure 2.8: *Coded Caching Scheme: Placement phase.*

Delivery phase: During the delivery phase, when the users start to make their requests for the contents, the desired contents that are not already present at the cache nodes are fetched from the server and transmitted to both the cache nodes using a common coded signal as shown in Fig. 2.9. The common coded signal has sub-files that are superimposed with each other and it depends on what the users request for during the delivery phase.

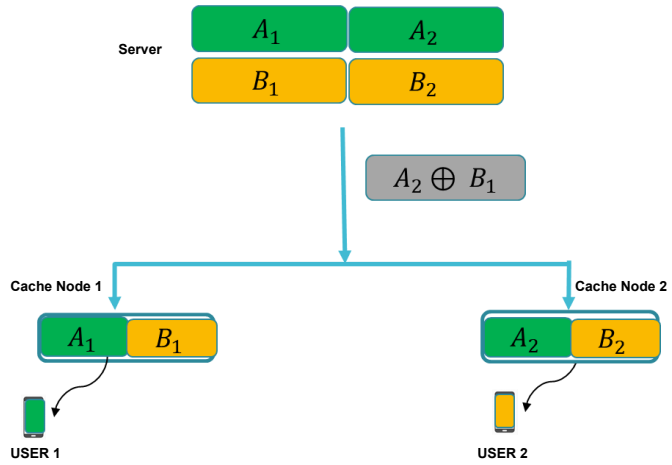

Figure 2.9: *Coded Caching Scheme: Delivery phase.*

For example, when User 1 requests for file A (green) and User 2 requests for file B (yellow) as shown in Fig. 2.9, a common coded signal comprising of sub-files that are not already present in the cache nodes is broadcast to both cache nodes simultaneously. In our example, the cache node 1 has A_1, B_1 from the placement phase and needs sub-file A_2 from the server to decode and retrieve the complete file A for the first user during the delivery phase. Similarly, cache node 2 has A_2, B_2 from the placement phase and needs sub-file B_1 from the server to decode and retrieve the complete file B for the second user during the delivery phase. Therefore, the common coded signal transmitted during the delivery phase would be $(A2 \oplus B1)$

where, $\oplus$ is the superposition operator. Tables 2.1, 2.2 and 2.3 show the common coded signals for different cases of user's requests and also the missing parts of the contents at each cache node.

User 1	User 2
A	A
A	B
B	A
B	B

Table 2.1: User Requests.

User 1	User 2
A_2	A_1
A_2	B_1
B_2	A_1
B_2	B_1

Table 2.2: Missing parts at the cache nodes.

Common coded signal	Backhaul load
$A_2 \oplus A_1$	0.5
$A_2 \oplus B_1$	0.5
$B_2 \oplus A_1$	0.5
$B_2 \oplus B_1$	0.5

Table 2.3: Common coded packets.

Since both users are simultaneously served from the common coded signal in just a single transmission, a multi-casting opportunity is exploited. By doing so, the backhaul load and congestion is reduced which, in turn reduces latencies in the network.

The worst-case peak-load obtained in this example using the CC scheme is 0.5 file, as opposed to the traditional/un-coded caching scheme with a worst-case backhaul load of 1 file (see Section 2.2.1.1). This is because, instead of fetching one whole file from the server as in the case of traditional caching scheme, only half a file is fetched during the delivery phase of the CC scheme.

The worst-case peak-load obtained in the CC scheme according to [16], [17], [18] is given by

$$R(M) = K\left(1 - \frac{M}{N}\right)\left(\frac{1}{1 + \frac{KM}{N}}\right) = 2\left(1 - \frac{1}{2}\right)\left(\frac{1}{2}\right) = 0.5. \tag{2.5}$$

2.2.2.2 Coded Caching: Example 2

Consider again another example of a network with $N = 3$ equally-popular files and $K = 3$ cache nodes. Assume that each user can cache up to $M = 1$ file during the placement phase and each user requests for a single file during the delivery phase. The three files are further divided into three sub-files of equal size which are concatenated as

$$\begin{aligned} \mathrm{A} - - > A_1, A_2, A_3, \\ \mathrm{B} - - > B_1, B_2, B_3, \\ \mathrm{C} - - > C_1, C_2, C_3. \end{aligned} \tag{2.6}$$

Placement phase: During the placement phase, the three cache nodes are filled with different portions or sub-files as shown in Fig. 2.10. That is, the cache nodes are filled with concatenated sub-files as follows:

$$\begin{aligned} \text{Cache node } 1 - - > (A_1, B_1, C_1), \\ \text{Cache node } 2 - - > (A_2, B_2, C_2), \\ \text{Cache node } 3 - - > (A_3, B_3, C_3). \end{aligned} \tag{2.7}$$

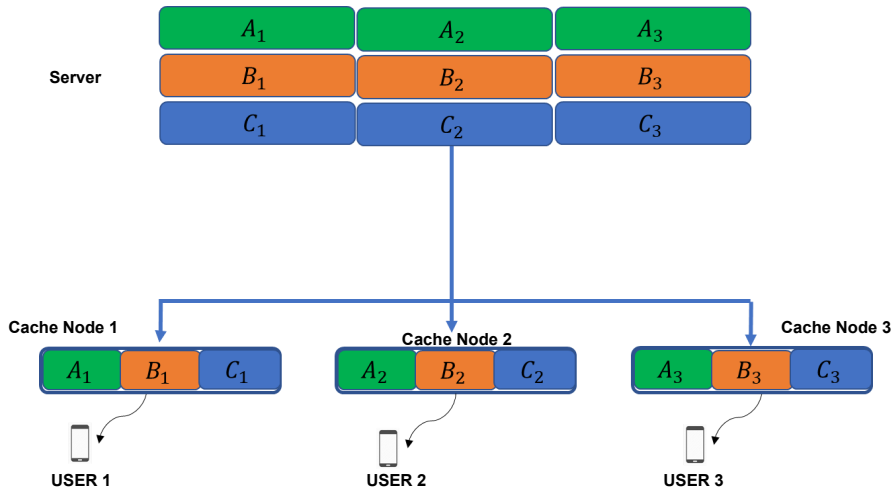

Figure 2.10: *Coded Caching Scheme example: Placement phase.*

Delivery Phase: During the delivery phase or high traffic period, when the users reveal their requests for the desired files, the cache nodes are checked to find any part of the requested files. The rest of the sub-files that are not already present in the cache node are broadcast to the cache nodes using common coded signals as shown in Fig. 2.11.

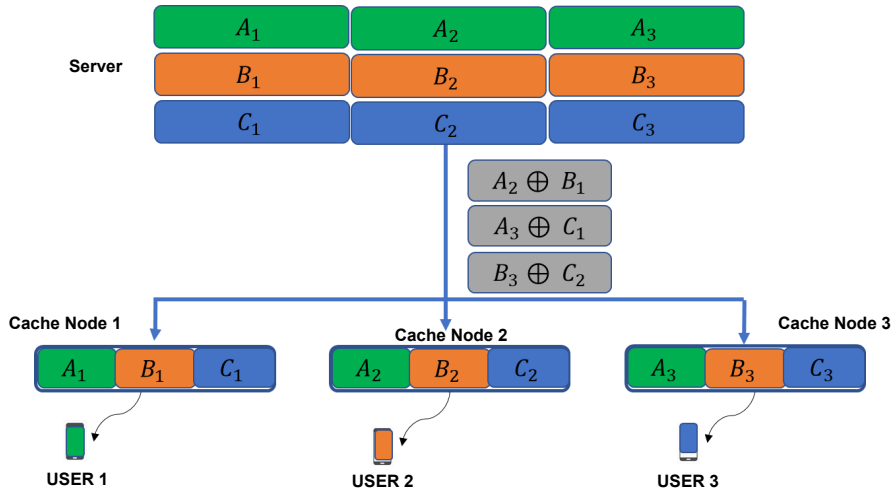

Figure 2.11: *Coded Caching Scheme example: Delivery phase.*

For example, when User 1 requests for file A (green), User 2 requests for file B (orange) and User 3 requests for file C (blue) as shown in Fig. 2.11, a set of common coded signals comprising of sub-files that are not already present in the cache nodes are broadcast to the cache nodes from the server sequentially. In our example, the cache node 1 has A_1, B_1 and C_1 from the placement phase and requires sub-files A_2 and A_3 from the server to decode and retrieve the complete file A for the first user during the delivery phase. Similarly, cache node 2 has A_2, B_2 and C_2 from the placement phase and requires sub-files B_1 and B_3 from the server to decode and retrieve the complete file B for the second user during the delivery phase. Finally, cache node 3 has A_3, B_3 and C_3 from the placement phase and requires sub-files C_1 and C_2 from the server to decode and retrieve the complete file C for the third user during the delivery phase. Therefore, the common coded signals transmitted during the delivery phase to cache nodes 1, 2 and 3 respectively would be:

Common coded signal broadcast to Cache nodes 1 and 2 : $(A2 \oplus B1)$.
Common coded signal broadcast to Cache nodes 1 and 3 : $(A3 \oplus C1)$.
Common coded signal broadcast to Cache nodes 2 and 3 : $(B3 \oplus C2)$.

The caches decode only the intended sub-files and the rest are either removed or treated as noise (see Section 3.3). The worst-case peak load for this example would then be 1 file since only 1 whole file i.e., three sub-files of size $\frac{1}{3}$ of a file is fetched from the server in the form of a set of common coded signals. The worst-case peak load is calculated as:

$$R(M) = K\left(1 - \frac{M}{N}\right)\left(\frac{1}{1 + \frac{KM}{N}}\right) = 3\left(1 - \frac{1}{3}\right)\left(\frac{1}{2}\right) = 1. \tag{2.8}$$

2.2.2.3 Coded Caching: General Case

When a general case with a single server consisting of N files serving K cache nodes is considered, we assume that the number of cache nodes and users are not more

than the number of files in the server. Each cache node is said to store up to M files during the placement phase. During the delivery phase, it is assumed that each user requests for a single file and the server broadcasts the sub-files that are not already cached during the placement phase to the cache nodes.

Placement phase: During the placement phase, the N files in the server are symmetrically divided into sub-files of equal size such that each cache node is able to store $\frac{M}{N}$ portion of each file that is present in the server. This placement/caching of sub-files in the CC scheme is based on the strategy proposed by Maddah-Ali and Niesen in [16], which is popularly known as the Maddah-Ali-Niesen (MAN) scheme. Therefore, each file is split into $\binom{K}{t}$ sub-files of equal size, where $t = \frac{KM}{N}$ [27]. Using this strategy, every cache node can cache up to $\frac{K-1}{t-1}$ sub-files out of an entire $\frac{K}{t}$ number of sub-files. Furthermore, these sub-files can be transmitted towards the intended cache-nodes using separate narrow-beams, which is discussed more in the upcoming sections.

Delivery phase: Once the users reveal their requests for the desired contents during the delivery phase, the server broadcasts common coded signals comprising of a superposition of sub-files, to serve the users with their intended requests. This common coded signal depends on the actual requests from the users as mentioned in Section 2.2.2.1. Because of the placement strategy followed according to the MAN scheme, the server can serve up to $(t+1)$ users by multicasting the common-coded signal to this set of users simultaneously in the network.

2.2.2.3.1 Performance Metrics In the literature, without considering the wireless network between the server and the cache nodes, different metrics are used to evaluate the performance of caching methods. Two main metrics are as follows:

- ***Worst-case peak load:*** The worst-case peak load, as defined in 2.2.2.1 for the CC scheme, can be calculated as [16]:

$$R(M) = K\left(1 - \frac{M}{N}\right)\left(\frac{1}{1 + \frac{KM}{N}}\right), \tag{2.9}$$

  where $\left(1 - \frac{M}{N}\right)$ is known as the *local caching gain* and $\left(\frac{1}{1+\frac{KM}{N}}\right)$ is known as the *coding gain*. The total gain obtained in the network that uses CC scheme is known as the *global caching gain* and is dependent on the aggregate of all caches in the network.

- ***Network throughput:*** The throughput of the network is defined as the amount of information or contents that is correctly delivered to the users per unit time. According to the MAN scheme, by considering a perfect channel between the server and the caches, the throughput is calculated as:

$$\text{throughput} = \left(\frac{1 + \left(\frac{KM}{N}\right)}{1 - \frac{M}{N}}\right). \tag{2.10}$$

2.3 Beamforming

5G may operate at millimeter wavelengths and the network is exposed to interferences like multipath effects, scattering or diffraction very easily. With the increase in the number of 5G networks and smartphones, the technique of beamforming plays a very crucial role [28].

Beamforming is a widely used technique in the signal processing, radar, biomedical industry and most importantly in communications. In communication systems, the beamforming technique is used to steer and focus the radiating radio frequency (RF) signals towards the intended users or a set of users. Instead of broadcasting a specific signal to all users that are within the range, beamforming uses multiple antennas to steer and send out the signal to, e.g., a single user such as a mobile device, laptop or a tablet which results in better connectivity and reliable data transfer [29], [30], [31].

The signal transmitted from a single antenna is nothing but electromagnetic waves. These electromagnetic waves, when radiated, tend to spread out in all directions until they encounter an interface like a physical object or a barrier like a wall. Therefore, to focus such waves or signal in a particular direction, multiple antennas are used. These multiple antennas are closely spaced to each other and broadcast the same signal at slightly different times [32]. This focused transmission of the signals can be achieved by choosing appropriate phase and power values at each stream of different antennas where the superposition of signals each with separate directivity is done. This adjusting or choosing of phase and power values determines the best path the signal should take to reach the intended user. That is, beamforming shapes the RF beam as it propagates in the physical space [29].

2.3.1 Types of Beamforming

Depending on the type or implementation, beamforming can be categorised into three types as follows:

1. **Analog Beamforming:** In early years, beamforming had an analog solution with fixed phase shifters to create beams at a single frequency [33]. Later, many phase shifters were incorporated in a switching architecture to produce multiple beam patterns. In the end, phase shifters that were adjustable were incorporated at each antenna element so the beams could travel in any direction. Figure 2.12 shows the basic architecture of the analog beamforming technique.

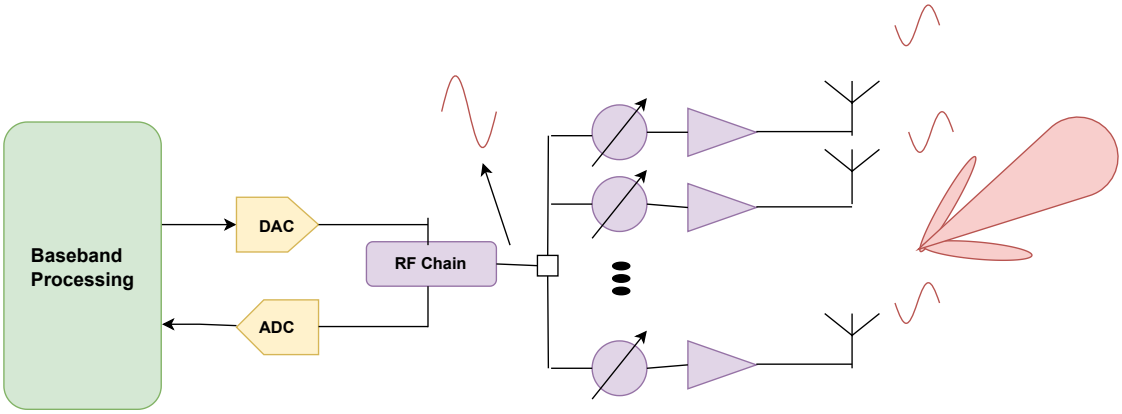

Figure 2.12: *Analog Beamforming.*

On the transmitter side, as shown in Fig. 2.12, the digital domain generates the baseband signal that is converted into an analog signal using the digital to analog converter (DAC). The signal is then up-converted and passed to the analog beamforming network through a splitter. In the beamforming network, weights are applied to the incoming signals by the phase shifters that are present at each antenna element [33]. The receiving side performs the opposite functions like phase shifting, combining and down-converting the received signal to a baseband frequency which is then converted to digital domain samples using an analog to digital converter (ADC).

The precoding or combining i.e., adjusting the amplitude and phase of the signals is done after the up-conversion in the transmitter or before the down-converstion in the receiver [28]. At the end of the analog beamforming technique, the same signal with different phases is produced at the end of each antenna element and transmitted into a specific direction.

Benefit: The analog beamforming technique is cost efficient since only a single pair of ADC and DAC and only one RF chain are used [33]. Because of this, the power is also efficiently consumed. Note that, by power efficient it simply means that the baseband power consumption is low but, the overall link is not power efficient and as analog beamforming is less precise, the received SINR might be lower than that of digital beamforming.

Limitations: Even though analog beamforming is cost and power efficient at a first glance, it may have the following limitations:

- The signals may be prone to interference in the unwanted directions as nulls cannot be generated in particular directions during the transmission or reception of the signal.
- Multiple users cannot be satisfied, that is, multi-user MIMO cannot be implemented as the same signal with different phases is generated at the end of this method [34].
- The analog components used in this network can introduce distortions and losses.

- Analog beamforming cannot perform frequency-adaptive beamforming in frequency selective channels.

2. **Digital Beamforming:** To overcome the limitations of analog beamforming, the technique of digital beamforming can be used [35]. A simple architecure of digital beamforming is as shown in Fig. 2.13.

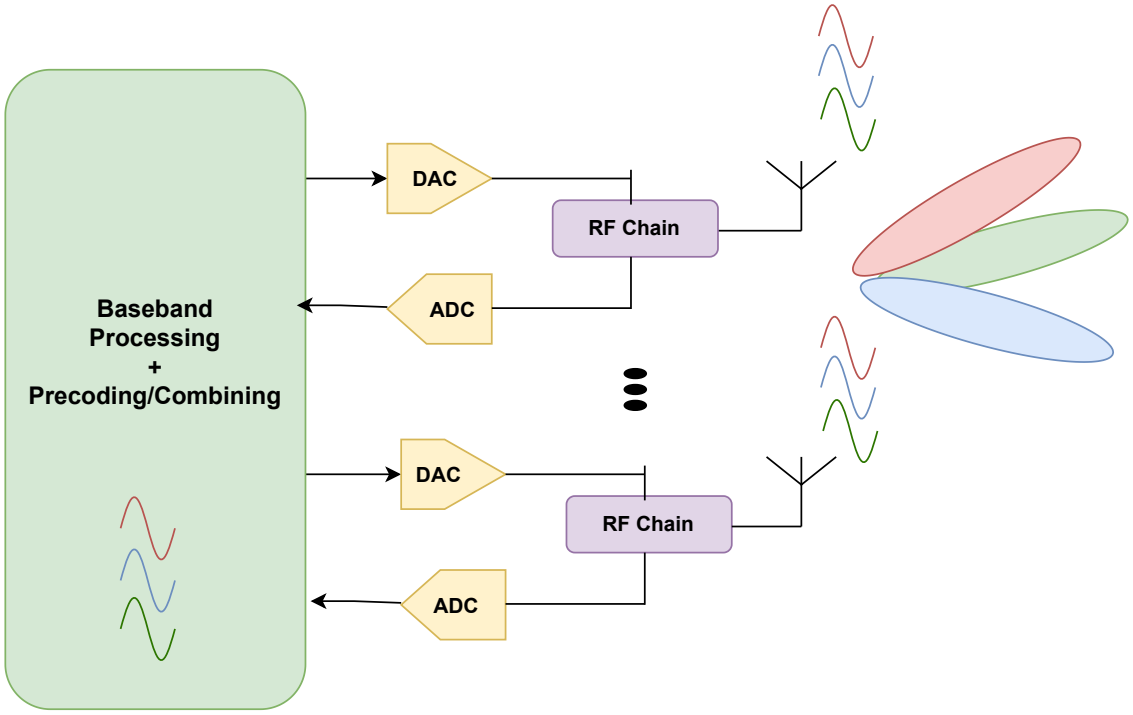

Figure 2.13: *Digital Beamforming.*

Instead of using only a single pair of DAC and ADC and one RF chain block like in the analog beamforming technique, digital beamforming uses a pair of DAC and ADC and an RF chain at each antenna element. The precoding or combining, that is adjusting the amplitude and phase of the signals is done before the up-conversion in the transmitter or after the down-conversion in the receiver [33].

Benefits:

- Null creation in the unwanted directions which reduces interferences in those directions and hence increases the signal strength in the desired direction. Hence, the coverage of a BS is also increased [36].
- Spatial multiplexing because of multiple spatial streams generated by digital beamforming [37]. Because of this, multiple users can communicate.

Limitation: The implementation cost and the power consumption in digital beamforming is high because many ADC, DAC components and RF chains are used in the architecture [36].

3. **Hybrid Beamforming:** The hybrid beamforming technique is a combination of analog and digital beamforming [35]. Analog beamforming method is less flexible but is cost and potentially power efficient, whereas the digital beamforming method is flexible but the cost and power consumption is high. Hence, a new technique, that is a compromise between the analog and digital bemaforming methods, called hybrid beamforming [38] was introduced. The hybrid beamforming method is mainly used for the mmWave frequency systems [39] and the simple architecture of it is as shown in Fig. 2.14.

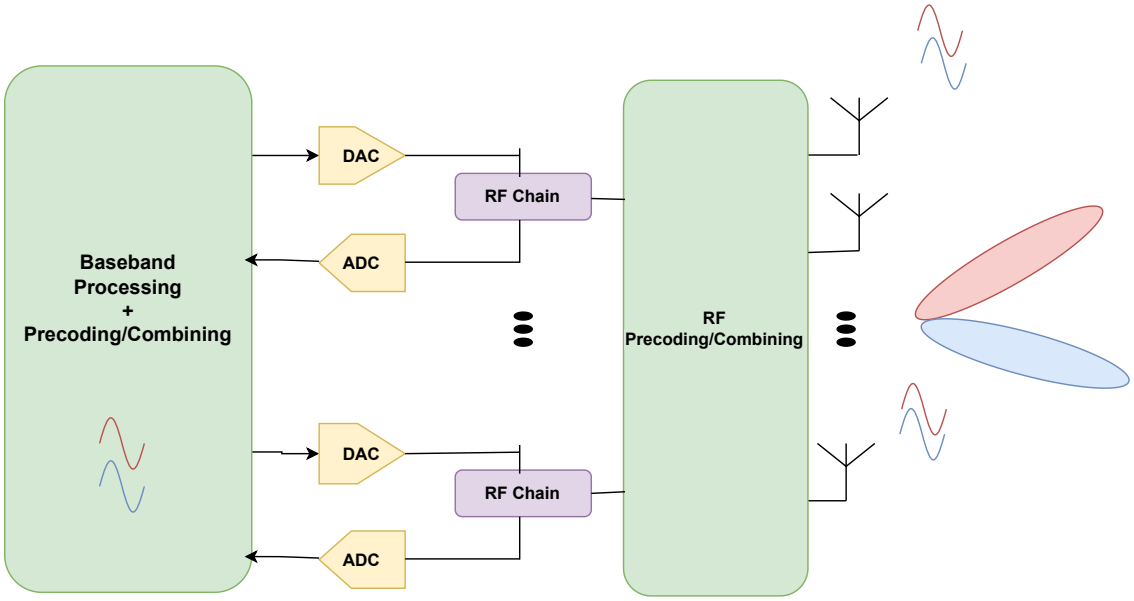

Figure 2.14: *Hybrid Beamforming.*

The precoding or combining, that is, adjusting the phase and amplitude of the signal at each stream, is done in both the analog as well as the digital domains. Because of this, components like ADC/DAC and RF chains can be fewer than the number of antennas. Hence the cost and power consumption remains efficient. This system is also flexible and can produce spatial streams on the order of the number of RF chains. Since the number of RF chains in hybrid beamforming is fewer than that of digital beamforming, the number of spatial streams generated are fewer than that of pure digital beamforming [33].

2.3.2 Benefits and Limitations of Beamforming

In general, incorporating the technique of beamforming in communication systems offers many advantages as well as disadvantages which are mentioned below:

Benefits:

- Provides high antenna gain and the signal strength is higher in the desired direction.
- Suppresses interference by creating nulls in the unwanted directions.

- Increases spectral efficiency which in turn improves the capacity in the network [34].
- Multiple users can be served by creating multiple spatial streams [37].
- Simple to implement analog beamforming as it is power and cost efficient [29].

Limitations and considerations:

- The cost can be high with the implementation of digital beamforming as it requires more resources like ADC/DAC pairs and RF chains.
- Along with the cost, the power consumption is also high with digital beamforming.
- With higher frequencies and more number of antennas, the implementation of massive MIMO and beamforming can be costly and complex [29].

2.4 Literature Review

The concept of caching is considered to be one of the important features in future wireless technologies [27] and is incorporated in various wireless network applications such as D2D [9]-[10], multi-hop networks [40], small cell networks [41], Coordinated Multi-Point (CoMP) [42] and cache-enabled helpers [43]. Caching is also seen in context-aware networks using edge/cloud computing which is investigated in [44]. The problems of edge caching is studied in [8] and information about CC is given in [45].

In [16], the concept of CC is proposed that achieves a global caching gain when a common coded signal is broadcast to different users requesting for different contents in the network. The extending works of [16] consider various structures like hierarchical CC [46], online CC [47] and multi-server scenarios [48]. All of the above mentioned works are from an information theoretical perspective.

When wireless network perspective is considered, various works that extend the concept of CC [16] in the high signal-to-noise-ratio (SNR) regime are observed. Different papers that consider the wireless interference channel, the effect of delayed channel state information at the transmitter (CSIT), and a wireless channel with mixed CSIT along with CC can be seen in [26], [49] and [50] respectively. Some works also consider interference management in wireless CC networks [51], rate splitting along with CC [25] and CC in wireless D2D networks [52].

When finite SNR regime is considered, CC for multiple-input single-output broadcast channel (MISO-BC) with rate splitting as the main idea can be seen in [53] and the same with multiple antennas is considered in [54]. The work in [55] considers CC along with zero-forcing (ZF) to further benefit from the spatial and global caching gains. The performance of CC in [54] and [55] can be further improved by carefully coding and designing beamformers, as can be seen in [56]. Further management of interferences is discussed in [27]. Also, CC along with multicast beamformers that are designed in the low SNR regime in [56] and [27] can be extended, to benefit from

the gains obtained from spatial multiplexing, global caching and by interference management because of transmitting the common coded signals in parallel.

3
Methods

3.1 Beamforming in Wireless Coded Caching Systems

Along with the CC scheme, the technique of beamforming can also be incorporated in the network to direct the packets or the sub-packets towards the intended users, so that the users can correctly decode the desired contents with better signal quality. As mentioned earlier in the CC scheme, the broadcasted common coded signals are visible to all intended users [56]. From one cache node perspective, the common coded signals contain both the sub-files that are desired and also the unwanted sub-files.

For example, consider a simple network having an L-antenna server with $N = 2$ equally-popular files (File A and File B) and $K = 2$ single-antenna cache nodes. Let us assume that each user can cache up to $M = 1$ file during the placement phase and each user requests for a single file during the delivery phase. The files in the server are divided into smaller sub-files as A_1, A_2 and B_1, B_2 respectively.

Placement phase: During the placement phase, when the cache nodes are filled with different portions of the files from the server, separate narrow beams can be used in different time slots to direct the transmission of the sub-packets towards their respective cache nodes as shown in Fig. 3.1.

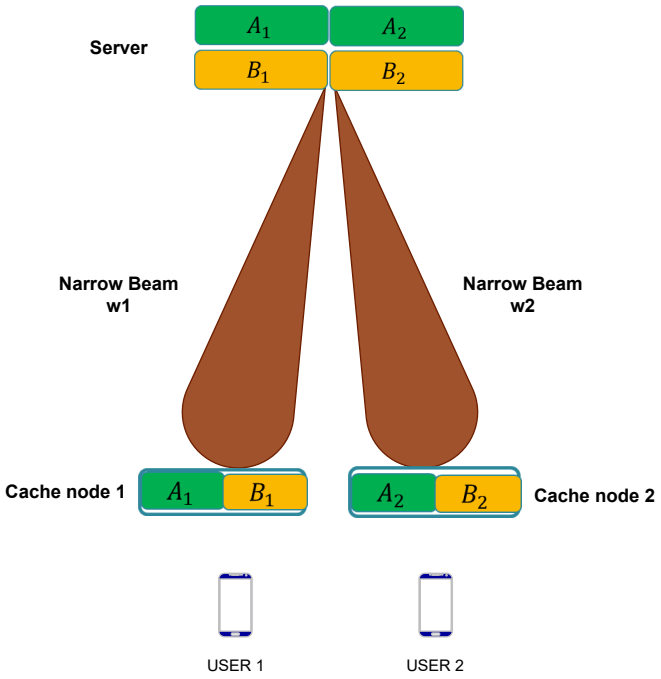

Figure 3.1: *Beamforming in Coded Caching: Placement phase where narrow beams are used in different time slots to serve the cache nodes.*

The signals transmitted to each cache node during the placement phase in the network is given by

$$Y_{C1} = \sqrt{P}\boldsymbol{h}_{\mathbf{1}}^{\mathrm{LT}}\boldsymbol{w}_{\mathbf{1}}[\tilde{A}_1(t)\tilde{B}_1(t)] + [Z_{C_{1,1}}(t)Z_{C_{1,2}}(t)], \tag{3.1}$$

$$Y_{C2} = \sqrt{P}\boldsymbol{h}_{\mathbf{2}}^{\mathrm{LT}}\boldsymbol{w}_{\mathbf{2}}[\tilde{A}_2(t)\tilde{B}_2(t)] + [Z_{C_{2,1}}(t)Z_{C_{2,2}}(t)], \tag{3.2}$$

where, P is the total transmit power of the server, $\boldsymbol{h}_{\boldsymbol{i}}^{\mathrm{LT}}$, $i = 1, 2$ is the channel realizations vector during the LT or placement phase for different cache nodes, $\boldsymbol{w}_{\boldsymbol{i_p}}$, $i = 1, 2$ are the beamforming vectors used to generate narrow beams for the placement of the sub-files at the respective cache nodes and $Z_{C_i,j}, i, j = 1, 2$, is the unit-variance additive Gaussian noise.

Delivery phase: During the peak traffic hours or the delivery phase, when the users reveal their requests for the desired contents, the common coded signal is broadcast to both cache nodes using a common wide beam as shown in Fig. 3.2.

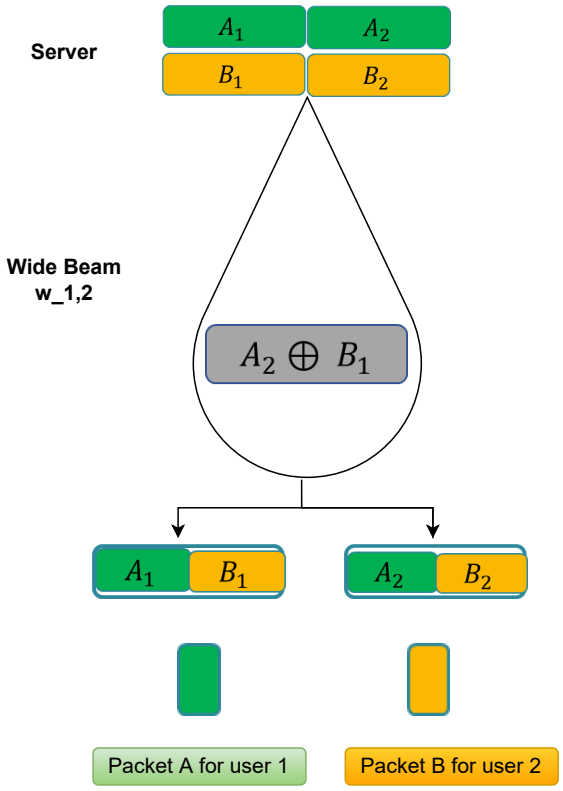

Figure 3.2: *Beamforming in Coded Caching: Delivery phase.*

The signals transmitted to each cache node during the placement phase in the network with $N = 2$ files and $K = 2$ users is given by

$$Y_{C1} = \sqrt{P}\boldsymbol{h}_{\mathbf{1}}^{\mathrm{HT}}\boldsymbol{w}_{\mathbf{1,2}}S(t) + Z_{C_{1,1}}(t), \tag{3.3}$$

$$Y_{C2} = \sqrt{P}\boldsymbol{h}_{\mathbf{2}}^{\mathrm{HT}}\boldsymbol{w}_{\mathbf{1,2}}S(t) + Z_{C_{2,2}}(t), \tag{3.4}$$

where, P is the total transmit power of the server, $\boldsymbol{h}_{\boldsymbol{i}}^{\mathrm{HT}}$, $i = 1, 2$ is the channel realizations during the high traffic or delivery phase for different users, $\boldsymbol{w}_{\mathbf{1,2}}$ is the beamforming vector used to generate a wide beam to transmit the common coded signal and $Z_{C_{i,j}}, i, j = 1, 2$, is the unit-variance additive Gaussian noise. The common coded signal in this example is given by

$$S(t) = \alpha\tilde{A}_2(t) + \sqrt{1-\alpha^2}\tilde{B}_1(t), \tag{3.5}$$

where, $\alpha \in [0, 1]$ is the power split parameter of the sub-files.

Once the common coded signal is broadcast to the cache nodes, each cache node tries to decode the intended content correctly using the already cached sub-file as well as the common coded signal. The details of the decoding methods will be explained in Section 3.3.

3.2 Optimization

In this thesis, the optimization of the beams for networks having beamforming along with CC scheme is done using the Genetic algorithm (GA). GA is an optimization technique that is based on natural selection and genetics. GA is part of a larger group of computation called evolutionary computation [30, 31, 57, 58, 59, 60, 61].

3.2.1 Genetic Algorithm

In optimization techniques like the GA, a set of possible solutions for the given formulation undergo modifications by, e.g., the mutation and recombination to produce a new set of children (solutions) and the process is repeated for many generations (iterations) [62]. Each individual/solution is assigned a fitness value (cost) according to the objective function and the fitter the individual, the higher are the chances for the individual to continue the process and produce more fitter individuals. This repeats for many generations until a stopping criterion is reached. Figure 3.3 shows the steps involved in GA.

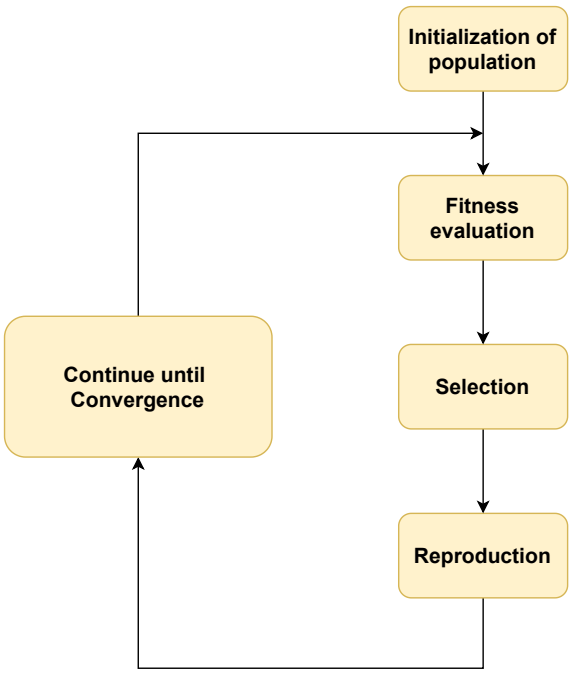

Figure 3.3: *Basic steps in GA.*

In this thesis, since we have beamforming only at the transmitter and not at the receivers, i.e., multiple transmit antennas at the transmitter and a single antenna at the receiver, the algorithm starts by creating sets of beams at the transmitter. These sets of beams are the sub-matrices chosen from the discrete fourier transform (DFT) based codebook. In every iteration, the best beam that is obtained depending on the performance metrics is chosen. This best beam that gives the best result is known

as the 'Queen' [58], [57]. As the performance metric for selecting the best solutions, we consider the maximum of the minimum signal-to-interference-noise-ratio (SINR) as observed by different cache nodes. Then, new sets of beams are generated around the Queen by choosing beams around the Queen and by randomly choosing a number of solutions from the DFT-based codebook for the remaining columns, so that the algorithm does not get stuck in the local minima [58]. At the end of a number of iterations that the designer chooses, the best beam that yields the best performance according to the mentioned performance metric is chosen.

3.2.2 Algorithm Description

Considering the example used in the previous section for beamforming in wireless CC system, the general steps used for the beam refinement using GA for a given time slot with instantaneous channel realizations are as shown below:

- **Step 1: Initialization:** Randomly pick a set of matrices of beams from the pre-defined DFT-based codebook.
- **Step 2: Selection:** For every precoding/beamforming matrix, calculate the objective metric (the minimum of the SINRs observed by different cache nodes in this case). Select the best beamforming matrix called as the 'Queen' that yields the best result for the objective metric.
- **Step 3: Queen:** Save the Queen.
- **Step 4:** Generate a new set of beamforming matrices by slightly modifying around the Queen.
- **Step 5:** The rest of the columns in the beamforming matrix are the columns randomly picked from the pre-defined DFT-based codebook.
- **Step 6:** Go back to Step 2 and iterate the procedure over a number of iterations set by the designer.
- **Step 7:** Return the Queen as the final beamforming solution in the considered time slot.

Note that the files in the server have information in them which are represented in nats. Suppose file A has K^{A} nats of information and file B has K^{B} nats, the information in the sub-files A_1, A_2, B_1, B_2 is divided as [63]

$$\begin{aligned} K^{\mathrm{A}} &= K_1^{\mathrm{A}} + K_2^{\mathrm{A}}, \\ K^{\mathrm{B}} &= K_1^{\mathrm{B}} + K_2^{\mathrm{B}}. \end{aligned} \tag{3.6}$$

Also, the code rates of the sub-packets are defined as

$$\begin{aligned} 2R^{\mathrm{A}} &= R_1^{\mathrm{A}} + R_2^{\mathrm{A}}, \\ 2R^{\mathrm{B}} &= R_1^{\mathrm{B}} + R_2^{\mathrm{B}}. \end{aligned} \tag{3.7}$$

where, R_i^{A} and R_i^{B}, $i = 1, 2$, are the rates of the sub-files. Here, (3.7) is based on (3.6) and the fact that the sub-files are of length L while the files are of length $2L$. Also, we define $R_i^{\mathrm{A}} = \frac{K_i^{\mathrm{A}}}{L}$, $R_i^{\mathrm{B}} = \frac{K_i^{\mathrm{B}}}{L}, i = 1, 2$, and $R^{\mathrm{A}} = \frac{K^{\mathrm{A}}}{2L}$, $R^{\mathrm{B}} = \frac{K^{\mathrm{B}}}{2L}$. Note that, in general, iterative algorithms may take time to converge. However,

- as we show in the simulations, our proposed GA converges with few iterations.
- Also, the server-cache node link is stationary, in which the beamforming update does not occur frequently.

It is interesting to note that in our method the beams of the placement and the delivery phases are optimized separately because in practice the placement and delivery phases occur in different periods and it is not possible to optimize them at the same time.

3.2.3 Optimization Problem

The network Successful Transmission Probability (STP) is defined as the average probability that the cache nodes decode their intended files correctly [63] and it is given by

$$\Pr\left(\mathrm{STP}\right) = \frac{1}{2}\left(\Pr\left(C1_{\mathrm{successful}}\right) + \Pr\left(C2_{\mathrm{successful}}\right)\right). \tag{3.8}$$

Here, C1 and C2 are cache nodes 1 and 2 respectively. Note that in (3.8) and the simulations, we concentrate on the cases with two cache nodes. However, it is straightforward to extend the results to the cases with multiple cache nodes. The final problem statement that would maximise the network STP by optimising the power split parameter α and the beamforming weights in $(3.1) - (3.4)$ is given as follows:

$$max\left(\mathrm{STP}\right) \tag{3.9}$$

such that the constraints mentioned below are met.

Constraints:

$$\begin{aligned} \alpha &\in [0, 1], \\ |w_1|^2 &\leq 1, \\ |w_2|^2 &\leq 1, \\ |w_{1,2}|^2 &\leq 1. \end{aligned} \tag{3.10}$$

The STP is calculated according to various decoding methods that is discussed in Section 3.3, w_i where $i = 1, 2$, are the beamforming weights for narrow beams during the placement phase, $w_{1,2}$ are the beamforming weights for the wide-beam during the delivery phase and P is the total transmit power. Note that the optimization of the power split parameter α is done through a simple exhaustive search while the beams are optimized by GA.

3.3 Decoding Techniques

The network STP can be maximized by properly choosing the power split parameter and also by optimizing the beamforming vectors. Depending on how the sub-files are cached or delivered, STP can be calculated for various decoding methods as follows.

3.3.1 Decoding Method 1: Joint Decoding during HT Period using Successive Interference Cancellation (SIC)

In this decoding Method 1, the sub-files received during the placement phase are cached and the intended sub-file is decoded using maximum ratio combining (MRC) and SIC during the delivery phase. Consider the following steps to decode the intended file.

1. Assume that User 1 requests for packet A and User 2 requests for packet B.
2. The idea is to first decode the unwanted sub-file from the received common coded signal using MRC between the received signals at the cache node during the placement and the delivery phases and subtract or remove it from the common coded signal using SIC. The decoding happens during the HT period or the delivery phase, where the intended sub-files are concatenated and the whole file is decoded in one shot.
3. Consider cache node 1. The signal received at cache node 1 during the delivery phase is given by Y_{C1} in (3.3).
4. The signals associated with sub-files $[\tilde{A}_1(t)\tilde{B}_1(t)]$ are already cached at cache node 1, but not decoded, during the placement phase.
5. The first step is to decode and remove $\tilde{B}_1(t)$ from the common coded signal given by Y_{C1} using MRC. Here, Y_{C1} in (3.3) and part of Y_{C1} in (3.1) associated with $\tilde{B}_1(t)$ are combined using MRC and decoded.
6. Next $\tilde{B}_1(t)$ is subtracted from the common coded signal that is, by using SIC to get an interference free signal which is given by
$$Y_{C1} = \alpha\sqrt{P}\boldsymbol{h}_{\mathbf{1}}^{\mathrm{HT}}\boldsymbol{w}_{\mathbf{1,2}}[\tilde{A}_2(t)] + Z_{C1,1}(t) \tag{3.11}$$
7. Lastly, we concatenate (3.11) and the part in (3.1) associated with $\tilde{A}_1(t)$ in (3.1) to decode the whole file A.
8. Similarly, cache node 2 follows the same procedure to decode the file B.

Let $g_1^{\mathrm{LT}} = |\boldsymbol{h}_{\mathbf{1}}^{\mathrm{LT}}\boldsymbol{w}_{\mathbf{1}_p}|^2$, $g_2^{\mathrm{LT}} = |\boldsymbol{h}_{\mathbf{2}}^{\mathrm{LT}}\boldsymbol{w}_{\mathbf{2}_p}|^{\mathbf{2}}$ be the channel gains obtained during the placement phase/LT period for cache node 1 and 2, respectively, and $g_1^{\mathrm{HT}} = |\boldsymbol{h}_{\mathbf{1}}^{\mathrm{HT}}\boldsymbol{w}_{\mathbf{1,2}}|^{\mathbf{2}}$, $g_2^{\mathrm{HT}} = |\boldsymbol{h}_{\mathbf{2}}^{\mathrm{HT}}\boldsymbol{w}_{\mathbf{1,2}}|^{\mathbf{2}}$ be the channel gains obtained during the delivery phase/HT period for cache node 1 and 2, respectively.

The network STP for the decoding Method 1 is calculated as follows [63]

$$\mathrm{STP} = \frac{1}{2}\left(\eta_1\gamma_1 + \eta_2\gamma_2\right), \tag{3.12}$$

where,

- Probability of successfully decoding $\tilde{B}_1(t)$ is given by

$$\eta_1 = \Pr\left(\log\left(1 + \left(\frac{(1-\alpha^2)\,Pg_1^{\mathrm{HT}}}{1+(\alpha^2 Pg_1^{\mathrm{HT}})}\right) + Pg_1^{\mathrm{LT}}\right) \geq R_1^B\right). \tag{3.13}$$

- Probability of successfully decoding file A after removing the interference $\tilde{B}_1(t)$ from the received common coded signal during the HT period is given by

$$\gamma_1 = \Pr\left(\log\left(1 + Pg_1^{\mathrm{LT}}\right) + \log\left(1 + \left(\alpha^2 Pg_1^{\mathrm{HT}}\right)\right) \geq 2R_A\right). \tag{3.14}$$

- Similarly, for cache node 2, the probability of successfully decoding $\tilde{A}_2(t)$ is given by

$$\eta_2 = \Pr\left(\log\left(1 + \left(\frac{(\alpha^2)\,Pg_2^{\mathrm{HT}}}{1+((1-\alpha^2)\,Pg_2^{\mathrm{HT}})}\right) + Pg_2^{\mathrm{LT}}\right) \geq R_2^A\right). \tag{3.15}$$

- Probability of successfully decoding $B(t)$ after removing the interference $\tilde{A}_2(t)$ from the received common coded signal during the HT period is given by

$$\gamma_2 = \Pr\left(\log\left(1 + Pg_2^{\mathrm{LT}}\right) + \log\left(1 + \left(\left(1-\alpha^2\right)Pg_2^{\mathrm{HT}}\right)\right) \geq 2R_B\right). \tag{3.16}$$

3.3.2 Decoding Method 2: Joint Decoding during HT Period without SIC

This method decodes the intended sub-files in one go by considering the interference as noise. That is, this decoding method does not use SIC. By doing so, the delay and complexity introduced in the implementation of decoding Method 1 can be reduced. Using decoding Method 2, (3.12) can be rearranged as [63]

$$\mathrm{STP} = \frac{1}{2}\left(\gamma_1 + \gamma_2\right), \tag{3.17}$$

where,

- The probability of decoding both the intended sub-file $\tilde{A}_2(t)$ received during the HT period and $\tilde{A}_1(t)$ that was already cached during the placement phase is given by

$$\gamma_1 = \Pr\left(\log\left(1 + Pg_1^{\mathrm{LT}}\right) + \log\left(1 + \left(\frac{(\alpha^2)\,Pg_1^{\mathrm{HT}}}{1+((1-\alpha^2)\,Pg_1^{\mathrm{HT}})}\right)\right) \geq 2R_A\right). \tag{3.18}$$

- The probability of decoding both the intended sub-file $\tilde{B}_1(t)$ received during the HT period and $\tilde{B}_2(t)$ already cached during the LT period is given by

$$\gamma_2 = \Pr\left(\log\left(1 + Pg_2^{\mathrm{LT}}\right) + \log\left(1 + \left(\frac{(1-\alpha^2)\,Pg_2^{\mathrm{HT}}}{1+((\alpha^2)\,Pg_2^{\mathrm{HT}})}\right)\right) \geq 2R_B\right). \tag{3.19}$$

3.3.3 Decoding Method 3: Separate Decoding with SIC

In decoding Method 3, the intended sub-files are decoded separately during the placement phase (LT) and during the delivery phase (HT). The network STP is then calculated as [63]

$$\mathrm{STP} = \frac{1}{2}\left(\eta_1\gamma_{11}\gamma_{12} + \eta_2\gamma_{21}\gamma_{22}\right), \tag{3.20}$$

where,

- η_1 and η_2 are same as in (3.13) and (3.15), respectively, and are the equations representing the probability of decoding $\tilde{B}_1(t)$ and $\tilde{A}_2(t)$ at cache nodes 1 and 2, respectively.
- The probability of cache node 1 decoding the intended sub-file $\tilde{A}_1(t)$ during the placement phase or LT period is given by

$$\gamma_{11} = \Pr\left(\log\left(1 + Pg_1^{\mathrm{LT}}\right) > R_1^A\right). \tag{3.21}$$

- The probability of cache node 1 decoding the intended sub-file $\tilde{A}_2(t)$ during the delivery phase or HT period after decoding the interfering sub-file $\tilde{B}_1(t)$ is given by

$$\gamma_{12} = \Pr\left(\log\left(1 + \alpha^2 Pg_1^{\mathrm{HT}}\right) > R_2^A\right). \tag{3.22}$$

- The probability of cache node 2 decoding the intended sub-file $\tilde{B}_1(t)$ during the delivery phase or HT period after decoding the interfering sub-file $\tilde{A}_2(t)$ is given by

$$\gamma_{21} = \Pr\left(\log\left(1 + \left(1 - \alpha^2\right) Pg_2^{\mathrm{HT}}\right) > R_1^B\right). \tag{3.23}$$

- The probability of cache node 2 decoding the intended sub-file $\tilde{B}_2(t)$ during the placement phase or LT time is given by

$$\gamma_{22} = \Pr\left(\log\left(1 + Pg_2^{\mathrm{LT}}+\right) > R_2^B\right). \tag{3.24}$$

3.3.4 Decoding Method 4: Separate Decoding without SIC

Decoding Method 4 considers decoding the intended sub-files separately, the same as in decoding Method 3 but, by considering the interference as noise. That is, the sub-files are decoded without using SIC. The network STP in this case is then calculated as [63]

$$\mathrm{STP} = \frac{1}{2}\left(\gamma_{11}\gamma_{12} + \gamma_{21}\gamma_{22}\right), \tag{3.25}$$

where,

- γ_{11} and γ_{22} are same as in (3.21) and (3.24), respectively.
- The probability of cache node 1 decoding the intended sub-file $\tilde{A}_2(t)$ by considering $\tilde{B}_1(t)$ interference as noise during the delivery phase or the HT period is given by

$$\gamma_{12} = \Pr\left(\log\left(1 + \left(\frac{\left(\alpha^2\right) Pg_1^{\mathrm{HT}}}{1 + \left(\left(1 - \alpha^2\right) Pg_1^{\mathrm{HT}}\right)}\right)\right) > R_2^A\right). \tag{3.26}$$

- The probability of cache node 2 decoding the intended sub-file $\tilde{B}_1(t)$ by considering $\tilde{A}_2(t)$ the interference as noise during the delivery phase or the HT period is given by

$$\gamma_{21} = \Pr\left(\log\left(1 + \left(\frac{(1-\alpha^2)\, P g_2^{\mathrm{HT}}}{1 + ((\alpha^2)\, P g_2^{\mathrm{HT}})}\right)\right) > R_1^B\right). \tag{3.27}$$

4

Simulation results

The simulation results are generated for a network with a single server consisting of multiple or a single transmit antennas serving two cache nodes. The channel considered in this thesis is the Rayleigh fading channel. The parameters considered for the simulation setup are as follows:

- $L = 10$, 16 or 32 transmit antennas at the server for a network that incorporates CC/ un-coded caching and beamforming and $L = 1$ for the case when the network incorporates CC/ un-coded caching without any beamforming.
- The total transmit power range is -10:10:60 dB.
- $K = 2$ cache nodes.
- Number of considered channel realizations is 15000.
- The number of GA iterations during for the placement and delivery phases is 150 iterations.
- The pre-defined considered date rates are [0.5:0.5:8] nats-per-channel-use (npcu) and we always consider $R^A = R^B$.

The STP and the throughput are calculated for networks incorporating CC and un-coded caching schemes with the above parameters. The same metrics are also calculated and compared for various decoding methods as mentioned in Section 3.3.

4.1 CC and Un-coded Caching without Beamforming

Figures 4.1 and 4.2 show the STP versus the transmit power in networks having un-coded caching and CC schemes without beamforming, respectively. The performance of both schemes is compared for various decoding methods as mentioned in Section 3.3. Here, we consider 2 cache nodes, data rate of 2 ncpu, transmit power of -10:10:50 dB, no beamforming and 1 antenna at the server.

In un-coded caching, only decoding Methods 2 and 4 are considered. This is because, in the un-coded caching scheme, since we are transmitting the signals in two separate narrow beams and at different time slots, there are no interferences between the signals and hence SIC is not needed. Thus, Methods 1 and 3, which follow SIC-based interference cancellation of the superimposed signals, are not of interest in the un-coded caching scheme. Whereas, in the CC scheme, there is interference and SIC may be considered. When the plot of network STP versus transmit power for decoding Methods 2 and 4 are compared in Fig. 4.1 and Fig. 4.2, network having un-coded caching scheme performs better than the network utilizing the CC scheme

, in terms of STP, because of no interference between the signals. However, this is at cost of extra transmissions and, consequently, lower throughput (see Figs. 4.7 and 4.8).

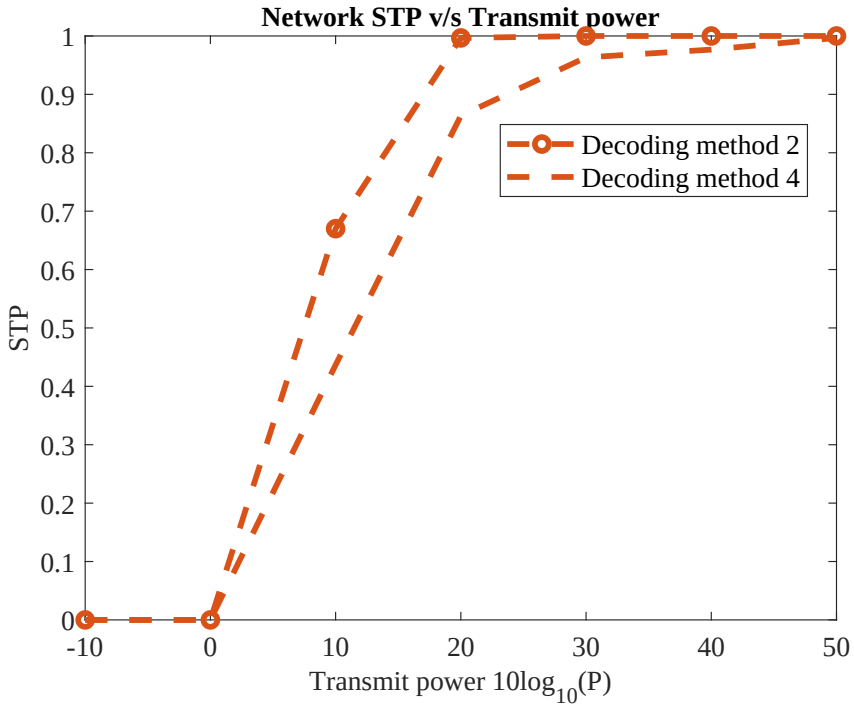

Figure 4.1: *Network STP v/s the transmit power in un-coded caching scheme without beamforming for a network with 2 cache nodes, data rate of 2 ncpu, transmit power of -10:10:50 dB and 1 antenna at the server.*

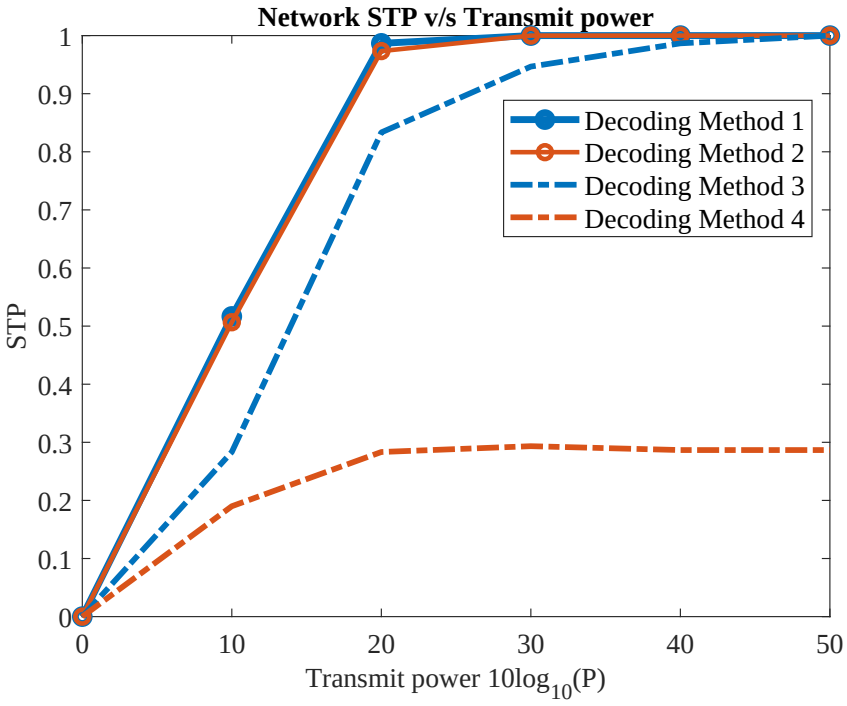

Figure 4.2: *Network STP v/s the transmit power in CC scheme without beamforming for a network with 2 cache nodes, data rate of 2 ncpu, transmit power of -10:10:50 dB and 1 antenna at the server.*

4.1.1 Beamforming with CC

When beamforming along with CC is incorporated in a network, GA-based optimization technique has been used to obtain the best beam during the placement and delivery phases according to the specified performance metric as mentioned in Section 3.2.
In Fig. 4.3, different examples of the convergence of the GA are demonstrated for different iterations of Algorithm 1. The figure demonstrates the minimum SINR observed by both cache nodes as the number of iterations increases. Here, we consider 2 cache nodes, transmit power of 60 dB, 32 antennas at the server and 150 GA iterations. As shown in Fig. 4.3, the GA converges with a few number of iterations. Also, the GA may temporally be trapped in a local minimum. However, due to Step 5 of the Algorithm 1, the GA can come out of the local minimum and converge towards the global optimum solution.

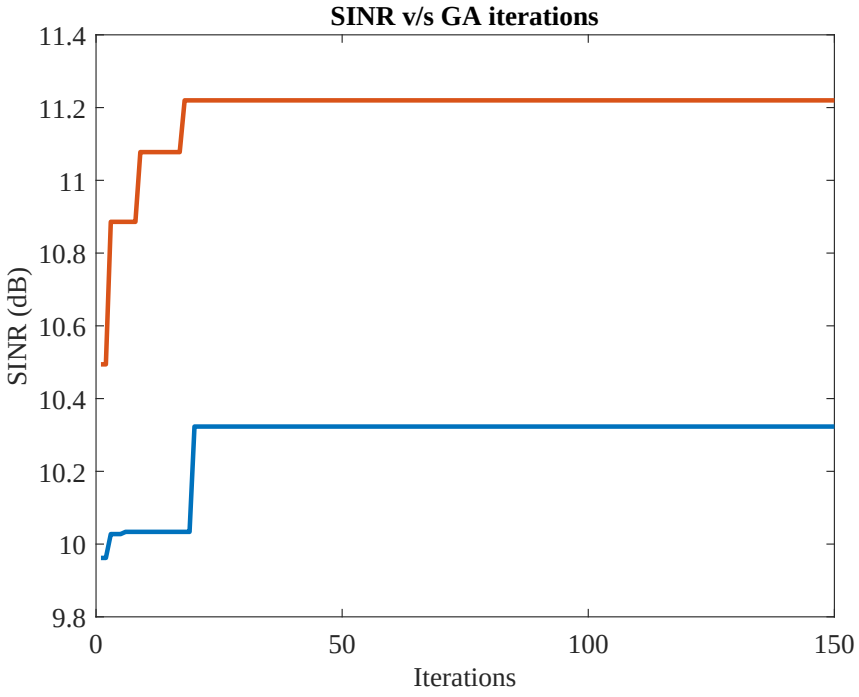

Figure 4.3: *Minimum SINR of the cache nodes v/s the GA iterations for a network with 2 cache nodes, transmit power of 60 dB, 32 antennas at the server and 150 GA iterations.*

Figure 4.4 shows the STP versus the transmit power for a network that incorporates CC scheme along with beamforming and also compares the results for various decoding methods. Here, we consider 2 cache nodes, transmit power of -10:10:40 dB, 32 antennas at the server and data rate of 2 ncpu. As observed from Fig. 4.4, we see that the performance of decoding Methods 1 and 2 are similar and that of decoding Method 1 is slightly better than that of decoding Method 2. It is also observed that there is a significant difference between the performances of decoding Method 3 and 4 when compared to decoding Methods 1 and 2. Note that the same point is observed in the cases with no beamforming, as shown in Fig. 4.2. This indicates that, compared to separate decoding, the joint decoding of the sub-files improves the network STP considerably. Note that joint decoding of Methods 1 and 2 is at the cost of additional decoding complexity. However, the drawback with Methods 3 and 4 is that, there the cache nodes may unnecessarily decode sub-files during the LT period which are not later requested by the users during the HT period. On the other hand, with proper beamforming and joint decoding, the effect of SIC-based interference cancellation on the network STP is negligible. This is the reason that the difference between the STP of Methods 1 and 2 is negligible. On the other hand, with separate decoding of the sub-files, interference cancellation becomes more important leading to considerably better STP in Method 3, compared to Method 4.

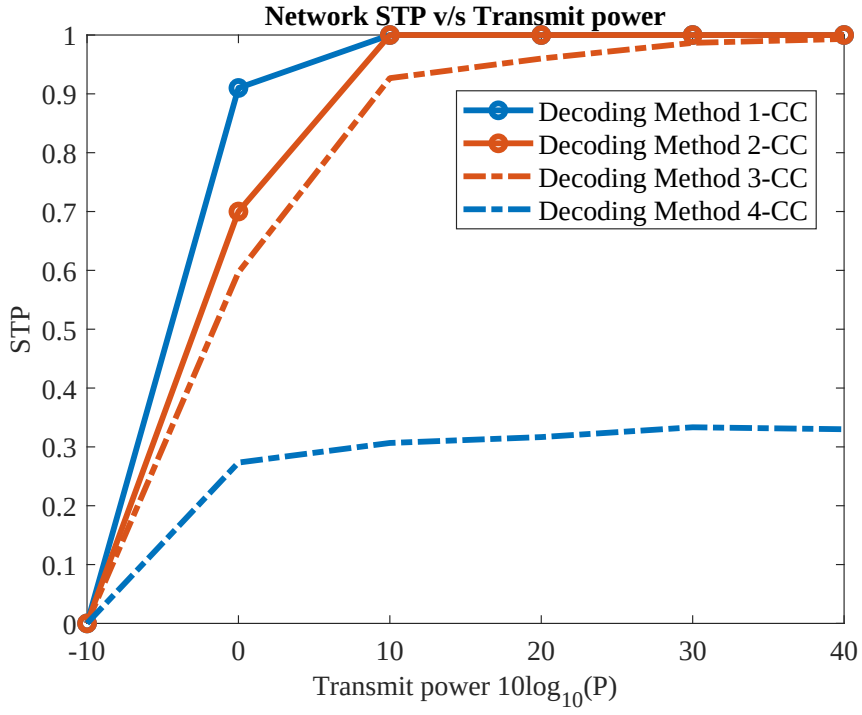

Figure 4.4: *Network STP v/s the transmit power in CC scheme with beamforming in a network with 2 cache nodes, transmit power of -10:10:40 dB, 32 antennas at the server and data rate of 2 ncpu.*

4.1.2 Comparison of Beamforming with CC and Un-coded Caching schemes

In this section, the performances of networks having CC and un-coded caching schemes along with beamforming are compared.
The STP is calculated and compared for various decoding methods as shown in Fig. 4.5. Here, we consider 2 cache nodes, transmit power of -10:10:40 dB, 32 antennas at the server and data rate of 2 ncpu. From Fig. 4.5, it can be immediately seen that, in terms of STP, the network incorporating the un-coded caching scheme along with beamforming performs better than the network having the CC scheme along with beamforming. For example, when decoding Method 2 is considered, the gap in the performance in terms of STP is very small. However, for the considered parameter settings of the figure, in Method 4, we see a significant gap between the STP of the networks with CC and un-coded caching schemes along with beamforming. This is intuitive because, as opposed to the CC scheme in which the signals are superimposed during the delivery phase at the cost of additional interference, in the un-coded schemes the sub-files are transferred in different time slots with no interference. However, as we show in the following, the implementation of un-coded scheme leads to considerable throughput drop, compared to the CC method (see Figs. 4.7 and 4.8).

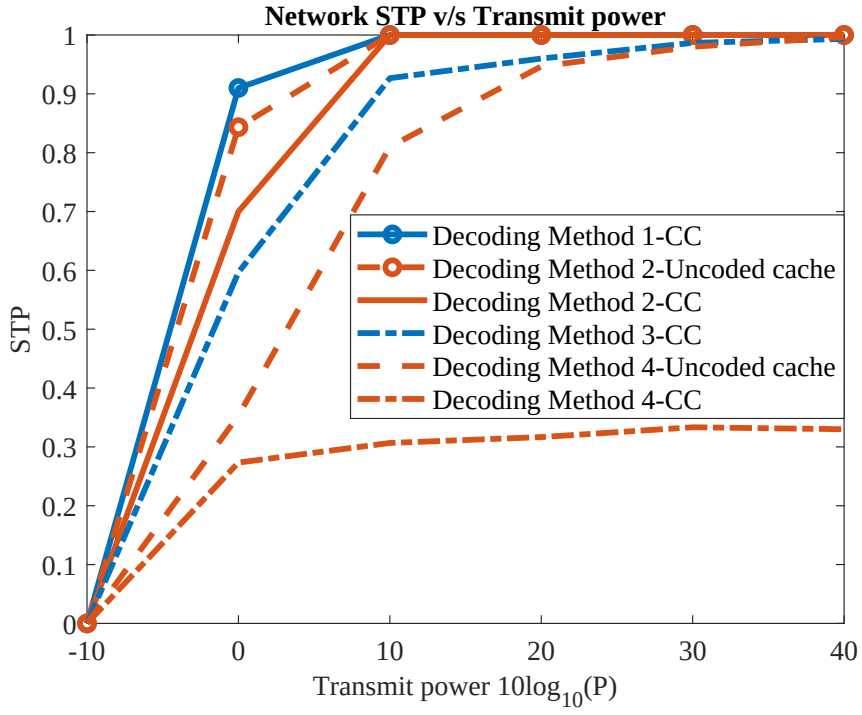

Figure 4.5: *Network STP v/s the transmit power in un-coded caching and CC schemes with beamforming for a network with 2 cache nodes, transmit power of -10:10:40 dB, 32 antennas at the server and data rate of 2 ncpu.*

Figure 4.6 shows a comparison of the network STP versus the data rate for both CC and un-coded caching incorporated networks for various decoding methods. Here, we consider 2 cache nodes, transmit power of 20 dB, 32 antennas at the server and data rates of [0.5:0.5:8] ncpu. Similar to Fig. 4.5, the performance (in terms of STP) of the network having un-coded caching scheme with beamforming is better than that of a network having CC scheme with beamforming. Also, there is a significant gap between the STPs of Methods 1 and 2 (joint decoding) and Methods 3 and 4 (separate decoding) before they all reach an STP value of 0. It can also be noticed from Fig. 4.6 that, Method 3 (separate decoding with SIC) tends to perform better than Method 2 (joint decoding without SIC) when the data date increases. Thus, interference cancellation is more important, than joint decoding at high data rates.

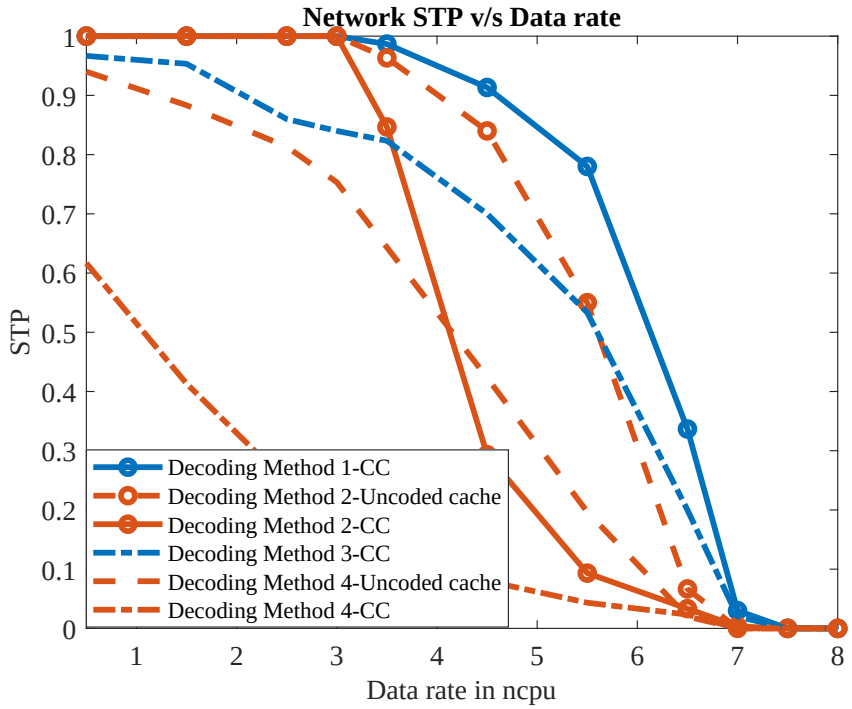

Figure 4.6: *Network STP v/s the data rate in un-coded caching and CC schemes for a network with 2 cache nodes, transmit power of 20 dB, 32 antennas at the server and data rate of [0.5:0.5:8] ncpu.*

Figure 4.7 shows the network throughput versus the number of transmit antennas at the server. Here, for the CC scheme, the throughput is defined as

$$\eta = R_1 \Pr(C1_{\text{successful}}) + R_2 \Pr(C2_{\text{successful}}), \tag{4.1}$$

where $R_i, i = 1, 2$, are the data rates for communications to cache node i. Also, $\Pr(C1_{\text{successful}})$ and $\Pr(C2_{\text{successful}})$ are the successful decoding probability in cache nodes 1 and 2, respectively, which for different Methods 1-4 are given in Section 3.3. On the other hand, for the un-coded caching methods, the throughput is given by

$$\eta = \frac{1}{2}(R_1 \Pr(\hat{C}1_{\text{successful}}) + R_2 \Pr(\hat{C}2_{\text{successful}})), \tag{4.2}$$

where $\Pr(\hat{C}1_{\text{successful}})$ and $\Pr(\hat{C}2_{\text{successful}})$ are the successful decoding probability in the un-coded caching methods. Note that the term $\frac{1}{2}$ in (4.2) comes from the fact that, in the un-coded caching scheme during the HT period, the signals are transmitted in two time slots, as opposed to the CC method with one-shot transmission of the superimposed signals.

Here, we consider 2 cache nodes, transmit power of -10:10:20 dB, $L = 10, 16, 32$ antennas at the server and a data rate of 2 ncpu and as observed from Fig. 4.7, the throughput of the network having CC scheme along with beamforming is considerably higher and better compared to the network incorporating un-coded caching

scheme with beamforming. It can also be seen that, for all the considered methods, greater number of antennas at the server improves the throughput in the network.

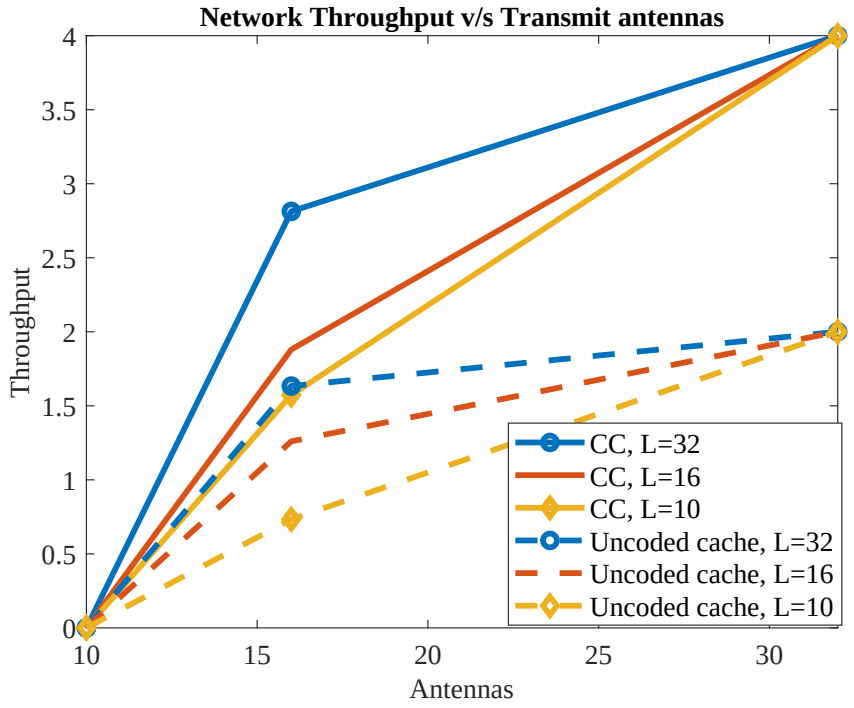

Figure 4.7: *Network throughput v/s the transmit antennas in CC and un-coded caching schemes with beamforming for a network with 2 cache nodes, transmit power of -10:10:40 dB, L = 10,16,32 antennas at the server and data rate of 2 ncpu.*

In Fig. 4.8, the throughput of the networks having CC and un-coded caching schemes along with beamforming with different numbers of DFT-based beams are compared. Here, we consider 2 cache nodes, transmit power of -10:10:40 dB, 32 antennas at the server and data rate of 2 ncpu. As observed from Fig. 4.8, the throughput of the network having CC with beamforming is significantly better than that of a network utilizing un-coded caching with beamforming. That is, although CC scheme reduces the STP slightly, compared to the un-coded caching method, the CC reduces the backhaul traffic at HT periods significantly.

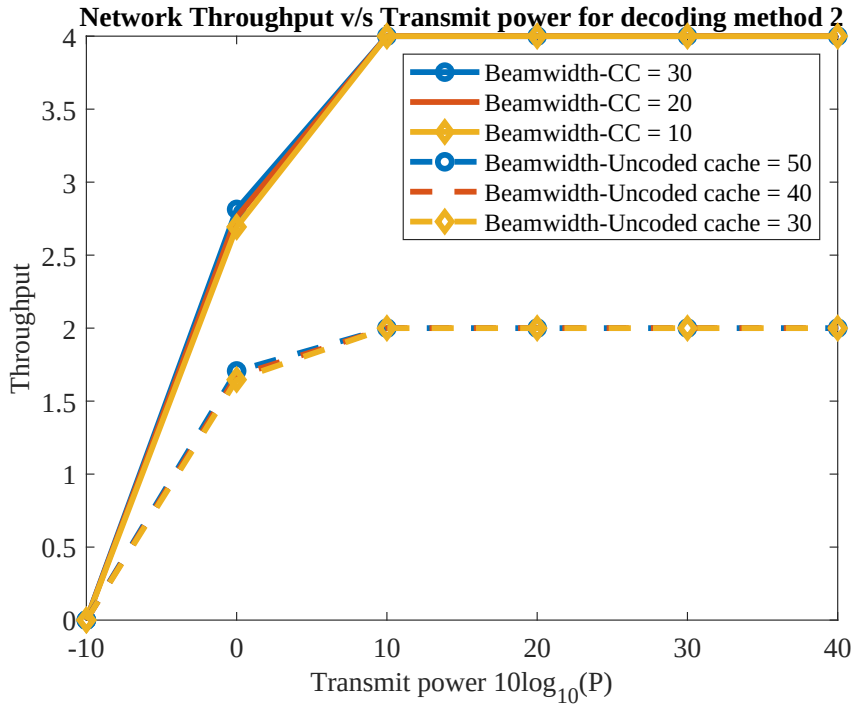

Figure 4.8: *Network throughput v/s the transmit power in CC and un-coded caching schemes with beamforming with different beamwidths for a network with 2 cache nodes, transmit power of -10:10:40 dB, 32 antennas at the server and data rate of 2 ncpu.*

4.1.3 Comparison of network having CC with and without beamforming

Figure 4.9 shows the comparison of network STP v/s the transmit power for Method 2 in networks incorporating CC and un-coded caching schemes with and without beamforming. Here, we consider 2 cache nodes, transmit power of -10:10:50 dB, 32 number of antennas at the server when beamforming is considered, 1 server antenna when no beamforming is considered and data rate of 2 ncpu. From Fig. 4.9, it is observed that the network having CC/un-coded caching schemes along with beamforming achieves a higher gain when compared to a network having CC/un-coded caching schemes without beamforming. It can also be observed from Fig. 4.9 that the network utilizing the un-coded caching scheme with or without beamforming performs slightly better than that of the network having CC with or without beamforming in terms of STP. This is because, the CC scheme overcomes the problem of delays in the network, but it has to deal with the interferences and different distances between the server and the cache node. In case of a network with un-coded caching scheme, the signals are more focused and do not have interferences and hence performs better in terms of STP, at the cost of throughput.

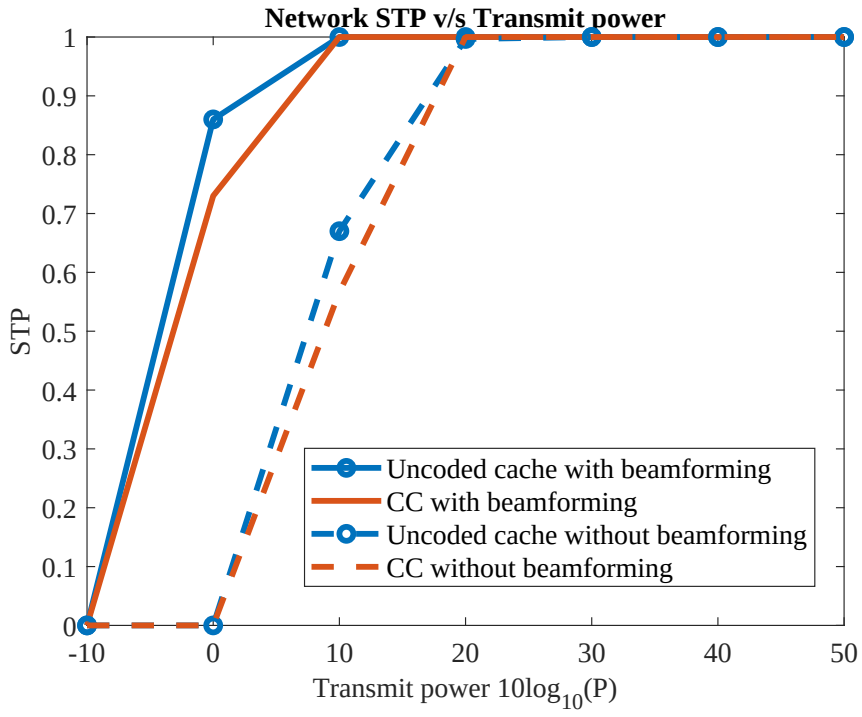

Figure 4.9: *Network STP v/s the transmit power in CC and uncoded caching schemes with and without beamforming for a network with 2 cache nodes, transmit power of -10:10:50 dB, 32 antennas at the server when beamforming is considered, 1 server antenna when no beamforming is considered and data rate of 2 ncpu.*

5
Conclusion

The impact of adaptive transmission and various decoding methods as well as different caching schemes on the networks with beamforming along with CC scheme was studied. As shown, using caching the backhaul traffic can be reduced and the network STP maximized by using beamforming along with CC. The interference and the delays in the network are reduced which results in better gains and network throughput. Furthermore, an efficient GA-based scheme for beamforming optimization in CC networks was developed, which shows promising results in reducing backhaul traffic.

6
Future Work

Although CC shows as a promising technique to reduce backhaul traffic, it has some practical issues. Most importantly, CC-based scheme should allow the cache nodes to decrypt the messages, which does not fit well with the end-to-end encryption requirements. There are few works on semi-encrypted CC, e.g., [64], [65], [66]. Other alternative technologies include introducing trusted representatives [67], [68], i.e., having "representatives" of the content provider at the network which have the users key and can decrypt the messages, or considering partial encrypted CC methods where each cache node can only decrypt its related messages but not those related to other caches, e.g., [69]. However, end-to-end encrypted CC is still an open problem to be solved.

Bibliography

[1] *Forecast number of mobile devices worldwide from 2020 to 2025 (in billions)**. [Online; accessed 2-June-2022]. URL: https://www.statista.com/statistics/245501/multiple-mobile-device-ownership-worldwide/.

[2] R. N. Mohammed H.Alsharif. "Evolution towards fifth generation (5G) wireless networks: Current trends and challenges in the deployment of millimetre wave, massive MIMO, and small cells". In: (Apr. 2017). DOI: 10.1007/s11235-016-0195-x.

[3] C. Madapatha, B. Makki, C. Fang, O. Teyeb, E. Dahlman, M.-S. Alouini, and T. Svensson. "On Integrated Access and Backhaul Networks: Current Status and Potentials". In: *IEEE open j. Commun. Soc.* 1 (Sept. 2020), pp. 1374–1389. DOI: 10.1109/OJCOMS.2020.3022529.

[4] C. Madapatha, B. Makki, A. Muhammad, E. Dahlman, M.-S. Alouini, and T. Svensson. "On Topology Optimization and Routing in Integrated Access and Backhaul Networks: A Genetic Algorithm-Based Approach". In: *IEEE open j. Commun. Soc.* 2 (Sept. 2021), pp. 2273–2291. DOI: 10.1109/OJCOMS.2021.3114669.

[5] *Scoring the Terabit/s Goal:Broadband Connectivity in 6G*. [Online; accessed 17-June-2022]. URL: https://arxiv.org/abs/2008.07220.

[6] *What is network densification?* [Online; accessed 2-June-2022]. URL: https://www.carritech.com/news/what-is-network-densification/.

[7] B. Romanous, N. Bitar, A. Imran, and H. Refai. "Network densification: Challenges and opportunities in enabling 5G". In: *Proc. IEEE CAMAD'2017*. Guildford, UK, Jan. 2015, pp. 129–134. DOI: 10.1109/CAMAD.2015.7390494.

[8] D. Liu, B. Chen, C. Yang, and A. F. Molisch. "Caching at the wireless edge: design aspects, challenges, and future directions". In: *IEEE Commun. Mag* 54.9 (Sept. 2016), pp. 22–28. DOI: 10.1109/MCOM.2016.7565183.

[9] M. Ji, G. Caire, and A. F. Molisch. "Wireless Device-to-Device Caching Networks: Basic Principles and System Performance". In: *IEEE J. Sel. Areas Commun.* 34.1 (Jan. 2016), pp. 176–189. DOI: 10.1109/JSAC.2015.2452672.

[10] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire. "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution". In: *IEEE Commun. Mag* 51.4 (Apr. 2013), pp. 142–149. DOI: 10.1109/MCOM.2013.6495773.

[11] X. Xu, Y. Xue, X. Li, L. Qi, and S. Wan. "A Computation Offloading Method for Edge Computing With Vehicle-to-Everything". In: *IEEE Access* 7 (Sept. 2019), pp. 131068–131077. DOI: 10.1109/ACCESS.2019.2940295.

[12] T. Zhang, S. Biswas, and T. Ratnarajah. "An Analysis on Wireless Edge Caching in In-Band Full-Duplex FR2-IAB Networks". In: *IEEE Access* 8 (Sept. 2020), pp. 164987–165002. DOI: 10.1109/ACCESS.2020.3022725.

[13] O. Teyeb, A. Muhammad, G. Mildh, E. Dahlman, F. Barac, and B. Makki. "Integrated Access Backhauled Networks". In: *Proc. IEEE VTC2019-Fall'2019*. Honolulu, HI, USA, Nov. 2019, pp. 1–5. DOI: 10.1109/VTCFall.2019.8891507.

[14] C. Fang, C. Madapatha, B. Makki, and T. Svensson. "Joint Scheduling and Throughput Maximization in Self-backhauled Millimeter Wave Cellular Networks". In: *Proc. ISWCS'2021*. Berlin,Germany, Oct. 2021, pp. 1–6. DOI: 10.1109/ISWCS49558.2021.9562232.

[15] O. P. Adare, H. Babbili, C. Madapatha, B. Makki, and T. Svensson. "Uplink Power Control in Integrated Access and Backhaul Networks". In: *Proc. IEEE DySPAN'2021*. Oct. 2021, pp. 163–168. DOI: 10.1109/DySPAN53946.2021.9677384.

[16] M. A. Maddah-Ali and U. Niesen. "Fundamental Limits of Caching". In: *IEEE Trans. Inf Theory* 60.5 (May 2014), pp. 2856–2867. DOI: 10.1109/TIT.2014.2306938.

[17] M. A. Maddah-Ali and U. Niesen. "Decentralized Coded Caching Attains Order-Optimal Memory-Rate Tradeoff". In: *EEE/ACM Trans. Netw.* 23.4 (Aug. 2015), pp. 1029–1040. DOI: 10.1109/TNET.2014.2317316.

[18] M. A. Maddah-Ali and U. Niesen. "Coding for caching: fundamental limits and practical challenges". In: *IEEE Commun. Mag* 54.8 (Aug. 2016), pp. 23–29. DOI: 10.1109/MCOM.2016.7537173.

[19] *Beamforming*. [Online; accessed 2-June-2022]. URL: https://en.wikipedia.org/wiki/Beamforming.

[20] A. McDonald. *Cisco: 79 % of world's mobile traffic to be video by 2022*. [Online; accessed 23-May-2022]. 2019. URL: https://www.digitaltveurope.com/2019/02/20/cisco-79-of-worlds-mobile-traffic-to-be-video-by-2022/.

[21] *IMT traffic estimates for the years 2020 to 2030*. [Online; accessed 1-June-2022]. URL: https://www.itu.int/dms_pub/itu-r/opb/rep/R-REP-M.2370-2015-PDF-E.pdf.

[22] *Caching for a Global Netflix*. [Online; accessed 17-June-2022]. URL: https://netflixtechblog.com/caching-for-a-global-netflix-7bcc457012f1.

[23] Q. Huang, K. Birman, R. Van Renesse, W. Lloyd, S. Kumar, and H. Li. "An analysis of Facebook photo caching". In: Nov. 2013, pp. 167–181. DOI: 10.1145/2517349.2522722.

[24] S. Mohajer, I. Bergel, and G. Caire. "Cooperative Wireless Mobile Caching: A Signal Processing Perspective". In: *IEEE Signal Process. Mag.* 37.2 (Mar. 2020), pp. 18–38. DOI: 10.1109/MSP.2019.2962507.

[25] E. Piovano, H. Joudeh, and B. Clerckx. "On coded caching in the overloaded MISO broadcast channel". In: *Proc. IEEE ISIT'2017*. Aachen, Germany, Aug. 2017, pp. 2795–2799. DOI: 10.1109/ISIT.2017.8007039.

[26] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr. "Fundamental Limits of Cache-Aided Interference Management". In: *IEEE Trans. Inf Theory* 63.5 (Feb. 2017), pp. 3092–3107. DOI: 10.1109/TIT.2017.2669942.

[27] A. Tölli, S. P. Shariatpanahi, J. Kaleva, and B. H. Khalaj. "Multi-Antenna Interference Management for Coded Caching". In: *IEEE Trans. Wireless Commun.* 19.3 (Mar. 2020), pp. 2091–2106. DOI: 10.1109/TWC.2019.2962686.

[28] *Analog vs. Digital Beamforming*. [Online; accessed 21-June-2022]. URL: https://blog.pasternack.com/antennas/analog-vs-digital-beamforming/.

[29] A. S. Gillis. *Beamforming*. [Online; accessed 30-May-2022]. URL: https://www.techtarget.com/searchnetworking/definition/beamforming.

[30] H. Guo, B. Makki, and T. Svensson. "A genetic algorithm-based beamforming approach for delay-constrained networks". In: *Proc. WiOpt'2017*. Paris, France, June 2017, pp. 1–7. DOI: 10.23919/WIOPT.2017.7959905.

[31] H. Guo, B. Makki, and T. Svensson. "A comparison of beam refinement algorithms for millimeter wave initial access". In: *Proc. IEEE PIMRC'2017*. Montreal, QC, Canada, Oct. 2017, pp. 1–7. DOI: 10.1109/PIMRC.2017.8292686.

[32] K. Shaw. *What is beamforming and how does it make wireless better?* [Online; accessed 30-May-2022]. URL: https://www.networkworld.com/article/3445039/beamforming-explained-how-it-makes-wireless-communication-faster.html.

[33] Q. Chaudhari. *What is the Difference between Analog, Digital and Hybrid Beamforming?* [Online; accessed 1-June-2022]. URL: https://wirelesspi.com/what-is-the-difference-between-analog-digital-and-hybrid-beamforming/.

[34] *Advantages of Analog Beamforming | disadvantages of Analog Beamforming*. [Online; accessed 21-June-2022]. URL: https://www.rfwireless-world.com/Terminology/Advantages-and-Disadvantages-of-Analog-Beamforming.html.

[35] *How to overcome the limitation of analog beamforming*. [Online; accessed 21-June-2022]. URL: https://ee-paper.com/how-to-overcome-the-limitation-of-analog-beamforming/.

[36] *Advantages of Beamforming | Disadvantages of Beamforming*. [Online; accessed 21-June-2022]. URL: https://www.rfwireless-world.com.

[37] *Analogue vs. Digital Beamforming*. [Online; accessed 21-June-2022]. URL: https://www.radartutorial.eu/06.antennas/Digital.

[38] C. Fang, B. Makki, J. Li, and T. Svensson. "Coordinated Hybrid Precoding for Energy-Efficient Millimeter Wave Systems". In: *Proc. IEEE SPAWC'2018*. Kalamata, Greece, June 2018, pp. 1–5. DOI: 10.1109/SPAWC.2018.8445887.

[39] C. Fang, B. Makki, J. Li, and T. Svensson. "Hybrid Precoding in Cooperative Millimeter Wave Networks". In: *IEEE Trans. Wireless Commun.* 20.8 (Mar. 2021), pp. 5373–5388. DOI: 10.1109/TWC.2021.3067636.

[40] S. Gitzenis, G. S. Paschos, and L. Tassiulas. "Asymptotic Laws for Joint Content Replication and Delivery in Wireless Networks". In: *IEEE Trans. Inf Theory* 59.5 (Dec. 2013), pp. 2760–2776. DOI: 10.1109/TIT.2012.2235905.

[41] E. Bastug, M. Bennis, and M. Debbah. "Living on the edge: The role of proactive caching in 5G wireless networks". In: *IEEE Commun. Mag* 52.8 (Aug. 2014), pp. 82–89. DOI: 10.1109/MCOM.2014.6871674.

[42] A. Liu and V. K. N. Lau. "Cache-Enabled Opportunistic Cooperative MIMO for Video Streaming in Wireless Systems". In: *IEEE Trans. Signal Process.* 62.2 (Nov. 2014), pp. 390–402. DOI: 10.1109/TSP.2013.2291211.

[43] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire. "FemtoCaching: Wireless Content Delivery Through Distributed Caching Help". In: *IEEE Trans. Inf Theory* 59.12 (Sept. 2013), pp. 8402–8413. DOI: 10.1109/TIT.2013.2281606.

[44] E. Zeydan, E. Bastug, M. Bennis, M. A. Kader, I. A. Karatepe, A. S. Er, and M. Debbah. "Big data caching for networking: moving from cloud to edge". In: *IEEE Commun. Mag* 54.9 (Sept. 2016), pp. 36–42. DOI: 10.1109/MCOM.2016.7565185.

[45] G. S. Paschos, G. Iosifidis, M. Tao, D. Towsley, and G. Caire. "The Role of Caching in Future Communication Systems and Networks". In: *IEEE J. Sel. Areas Commun.* 36.6 (June 2018), pp. 1111–1125. DOI: 10.1109/JSAC.2018.2844939.

[46] N. Karamchandani, U. Niesen, M. A. Maddah-Ali, and S. N. Diggavi. "Hierarchical Coded Caching". In: *IEEE Trans. Inf Theory* 62.6 (Apr. 2016), pp. 3212–3229. DOI: 10.1109/TIT.2016.2557804.

[47] R. Pedarsani, M. A. Maddah-Ali, and U. Niesen. "Online Coded Caching". In: *IEEE/ACM Trans. Netw.* 24.2 (Mar. 2016), pp. 836–845. DOI: 10.1109/TNET.2015.2394482.

[48] S. P. Shariatpanahi, S. A. Motahari, and B. H. Khalaj. "Multi-Server Coded Caching". In: *IEEE Trans. Inf Theory* 62.12 (Sept. 2016), pp. 7253–7271. DOI: 10.1109/TIT.2016.2614722.

[49] J. Zhang and P. Elia. "Fundamental Limits of Cache-Aided Wireless BC: Interplay of Coded-Caching and CSIT Feedback". In: *IEEE Trans. Inf Theory* 63.5 (Feb. 2017), pp. 3142–3160. DOI: 10.1109/TIT.2017.2674668.

[50] M. A. T. Nejad, S. P. Shariatpanahi, and B. H. Khalaj. "On storage allocation in cache-enabled interference channels with mixed CSIT". In: *Proc. IEEE ICC Workshops'2017*. Guildford, UK, July 2017, pp. 1177–1182. DOI: 10.1109/ICCW.2017.7962818.

[51] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr. "Cache-Aided Interference Management in Wireless Cellular Networks". In: *IEEE Trans. Commun.* 67.5 (Jan. 2019), pp. 3376–3387. DOI: 10.1109/TCOMM.2019.2893669.

[52] M. Ji, G. Caire, and A. F. Molisch. "Fundamental Limits of Caching in Wireless D2D Networks". In: *IEEE Trans. Inf Theory* 62.2 (Dec. 2016), pp. 849–869. DOI: 10.1109/TIT.2015.2504556.

[53] K.-H. Ngo, S. Yang, and M. Kobayashi. "Scalable Content Delivery With Coded Caching in Multi-Antenna Fading Channels". In: *IEEE Trans. Wireless Commun.* 17.1 (Nov. 2018), pp. 548–562. DOI: 10.1109/TWC.2017.2768361.

[54] S. P. Shariatpanahi, G. Caire, and B. H. Khalaj. "Multi-antenna coded caching". In: *Proc. IEEE ISIT'2017*. Aachen, Germany, Aug. 2017, pp. 2113–2117. DOI: 10.1109/ISIT.2017.8006902.

[55] S. P. Shariatpanahi, G. Caire, and B. Hossein Khalaj. "Physical-Layer Schemes for Wireless Coded Caching". In: *IEEE Trans. Inf Theory* 65.5 (Dec. 2019), pp. 2792–2807. DOI: 10.1109/TIT.2018.2888615.

[56] A. Tolli, S. P. Shariatpanahi, J. Kaleva, and B. Khalaj. "Multicast Beamformer Design for Coded Caching". In: *Proc. IEEE ISIT'2017*. Vail, CO, USA, Aug. 2018, pp. 1914–1918. DOI: 10.1109/ISIT.2018.8437354.

[57] B. Makki, T. Svensson, G. Cocco, T. de Cola, and S. Erl. "On the throughput of the return-link multi-beam satellite systems using genetic algorithm-based schedulers". In: *Proc. IEEE ICC'2015*. London, UK, June 2015, pp. 838–843. DOI: 10.1109/ICC.2015.7248426.

[58] H. Guo, B. Makki, and T. Svensson. "Genetic Algorithm-Based Beam Refinement for Initial Access in Millimeter Wave Mobile Networks". In: (June 2018). [Online; accessed 25-May-2022]. URL: https://doi.org/10.1155/2018/5817120.

[59] *Genetic algorithm*. [Online; accessed 6-June-2022]. URL: https://en.wikipedia.org/wiki/Genetic_algorithm.

[60] B. Makki, A. Ide, T. Svensson, T. Eriksson, and M.-S. Alouini. "A Genetic Algorithm-Based Antenna Selection Approach for Large-but-Finite MIMO Networks". In: *IEEE Trans. Veh. Technol* 66.7 (July 2017), pp. 6591–6595. DOI: 10.1109/TVT.2016.2646139.

[61] B. Makki, T. Svensson, and M.-S. Alouini. "On the Throughput of Large-but-Finite MIMO Networks Using Schedulers". In: *IEEE Trans. Wireless Commun.* 18.1 (Jan. 2019), pp. 152–166. DOI: 10.1109/TWC.2018.2878252.

[62] *Genetic Algorithms - Introduction*. [Online; accessed 6-June-2022]. URL: https://www.tutorialspoint.com/genetic_algorithms/genetic_algorithms_introduction.htm.

[63] B. Makki and M.-S. Alouini. "Coded-Caching Using Adaptive Transmission". In: *IEEE Trans. Wireless Commun.* 10.10 (Oct. 2021), pp. 2160–2164. DOI: 10.1109/LWC.2021.3095294.

[64] A. Araldo, G. Dán, and D. Rossi. "Caching Encrypted Content Via Stochastic Cache Partitioning". In: *EEE/ACM Trans. Netw.* 26.1 (Jan. 2018), pp. 548–561. DOI: 10.1109/TNET.2018.2793892.

[65] X. Yuan, X. Wang, J. Wang, Y. Chu, C. Wang, J. Wang, M.-J. Montpetit, and S. Liu. "Enabling Secure and Efficient Video Delivery Through Encrypted In-Network Caching". In: *IEEE J. Sel. Areas Commun.* 34.8 (June 2016), pp. 2077–2090. DOI: 10.1109/JSAC.2016.2577301.

[66] J. Leguay, G. S. Paschos, E. A. Quaglia, and B. Smyth. "CryptoCache: Network caching with confidentiality". In: *Proc. IEEE ICC'2017*. Paris, France, May 2017, pp. 1–6. DOI: 10.1109/ICC.2017.7996866.

[67] G. Paschos, E. Bastug, I. Land, G. Caire, and M. Debbah. "Wireless caching: technical misconceptions and business barriers". In: *IEEE Commun. Mag* 54.8 (Aug. 2016), pp. 16–22. DOI: 10.1109/MCOM.2016.7537172.

[68] J. Roberts and N. Sbihi. "Exploring the memory-bandwidth tradeoff in an information-centric network". In: *Proc. ITC'2013*. Paris, France, Nov. 2013, pp. 1–9. DOI: 10.1109/ITC.2013.6662936.

[69] V. Ravindrakumar, P. Panda, N. Karamchandani, and V. M. Prabhakaran. "Private Coded Caching". In: *IEEE Trans. Inf. Forensics Secur.* 13.3 (Oct. 2018), pp. 685–694. DOI: 10.1109/TIFS.2017.2765503.

DEPARTMENT OF ELECTRICAL ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY