UNIVERSITÀ DI PISA

Università degli Studi di Pisa

FACOLTÀ DI INGENGERIA
CORSO DI LAUREA MAGISTRALE IN
INGEGNERIA DELLE TELECOMUNICAZIONI

CONTENT DELIVERY OPTIMIZATION FOR HYBRID SATELLITE NETWORKS

Relatori:
Prof. Luca Sanguinetti
DII, Universitá di Pisa

Candidato:
Marilena Fittipaldi

Prof. Björn Ottersten
SnT, Université du Luxembourg

Correlatore:
Dr. Symeon Chatzinotas
SnT, Université du Luxembourg

Anno Accademico 2015-2016

Contents

List of Figures

List of Tables

Abstract

Data traffic is delivered in nowadays networks to users by means of content delivery networks (CDNs). The main idea behind current CDNs is to store contents as close as possible, at the edge of the network, to the end users according to files' popularities. To place contents at a single hop from users, the so-called edge base stations (E-BSs), equipped with data storage, can establish links with very low latency, due to their vicinity to the users, thus increasing the quality of experience (QoE). In order to fill the caches in an intelligent manner, E-BSs take into account local users' requests by storing files according to popularity distribution. In this way, most locally popular contents can be efficiently served. This approach, however, doesn't consider local users behaviours are influenced also by global trends. Hybrid networks, employing a satellite overlay over a preexisting CDN, can introduce important improvements. Despite local popularity is computed taking into account local users requests, the satellite, having a global view of the network, can extract global popularity trends of the users. The satellite sends simultaneously to all edge caches the most globally popular contents to store that will be most likely required in the near future. The challenge is to find an optimal placement algorithm that, using both local and global distributions, can achieve the best performance in terms of QoE, latency and bandwidth consumption.

This work investigates different offline placement algorithms. Their performance is evaluated based on the hit ratio, which is defined as the ratio between number of requests in-cache and total number of users requests, and the time placement, intended as the time needed to fill caches (using broadcast, in satellite, or unicast, in terrestrial, transmission). The local-only and the global-only implemented placements store most popular contents according to, respectively, local and global popularity distributions. They are used as baseline solutions to evaluate the proposed hybrid placement. We aim to show the hybrid placement algorithm can achieve better performance if compared to the non-hybrid approaches, in terms of users QoE and network resource allocation.

Chapter 1

Terrestrial Content Delivery Networks

In recent years, a rapid growth of mobile users, mainly due to the advent of smartphones, is occurring. Global IP traffic has increased eight times between 2013 and 2016, an unprecedented rate, which is expected to hold per next years, bringing new challenges to network communications [2]. Internet traffic today is made by web objects (text, graphics, URLs and scripts), downloadable objects (media files, software, documents), applications (e-commerce, portals), live streaming media, on demand streaming media, and social networks. Video on Demand (VoD), is seen as one of the most demanding services. In fact, two-thirds of all requested traffic is made by VoD [1] right now and it is predicted to grow up to 70% by 2021, due to the user preferences are going towards video based applications, such as YouTube and NetFlix [3]. Network operators themselves cannot afford this huge amount of traffic. The backhaul link, connecting the servers to the core network, is the portion of the network most suffering by this traffic load as it responsable to conveys both the user requests to the servers and the desired contents to the users. Indeed when the amount of requested data is significant, the bachkaul becomes congested producing delay in the delivery and decreasing users' QoE (Quality of Experience) and system performance [4]. A possible solution to overcome problem is to use Content Delivery Networks (CDNs), which are systems to distribute contents among geographically servers, equipped with caches to store appropriate contents [1]. Intermediate nodes are an intuitive and intelligent way to bypass the backhaul, avoiding congestion of the network. In fact, to face the traffic requirements, advanced CNDs aim to increase capillarity, storing contents at the edge of the network closer to end users. The best solution to minimize the latency is to store contents at one hop from users, in the so-called edge base stations (E-BSs).

Caching at the edge base stations is a promising approach to reduce latency and backhauling load, in particular, the more distributed and efficient are located the edge base stations and the way to store contents, the better the service offered to clients. In order to do this, a **user content oriented network** is needed to make a network-layer aware of the requested content. Consequently, an efficient network has to focus on users behaviour as well as on users' distribution [5] to discover **who is asking for what, where and how often**, which essentially means **to model contents' popularity**. Each edge base station, in fact, can calculate files popularity counting received requests from its users. A local distribution is available and it can be used to fill edge cache in an intelligent way, avoiding data requests passing through backhaul links. This brings network improvements and increased users' QoE.

1.1 Terrestrial Content Delivery Network architecture

A Content Delivery Network is a network made by servers, equipped with data capabilities which are used to efficiently store and distribute data, mainly large dimensions files, like videos, all over the world.
In order to delivery contents, two different network approaches exist: the centralized and the decentralized, respectively shown in Figure 1.1a, 1.1b. In the centralized network approach, a central single server has to oversee all contents, achieving significant performance degradation when video on demand and streaming TV are the main required traffic [6]. Unlike, CDNs create an embedded content-oriented network, in which servers are distributed geographically. Starting from the Internet end-to-end principle, where an information pass through two reference nodes, the host and the end user, the main idea is to replicate this paradigm using distributed server nodes to whom users can refer, asking for contents, as shown in Figure 1.1.

(a) (b)

Figure 1.1: Network architectures with centralized and decentralized approach

A schematic representation of CDN architecture is shown in Figure 1.2:

Figure 1.2: Content Delivery Network

At the center of the system, the core network collects all available contents for a specific provider. Other distributed nodes store in their local caches some original files' perfect copies, and distribute them in order to make available to users, more easily. When a user needs a content, a request is sent to its reference server. If the requested file is stored, it can be directly sent through the terrestrial link otherwise the request has to pass through the network, wasting resources and increasing the latency. Since contents are not required with the same frequency, a server receives different requests rate for each of them. The requests follow a known popularity profile, as explained in Section 1.3

The challenge is to understand where caches have to be located and which content

is better to store there. Caching contents in different parts of the network, in fact, is the real cornerstone of this system, routing the incoming request to the node which is considered the best for the delivery [7]. Splitting contents in distributed servers, as done in the first approach with CDNs, has some adavantages, if compared to the centralized approach, like reduction of request response time, network bandwidth consumption and the server load [8]. Nevertheless, traffic requires a more variegate distribution of contents to avoid the use of the backhaul link.
In order to optimize the delivery process, ensuring the core network has not traffic peaks and avoiding congestions, CDNs aim to increase network capillarity [7], [9], saving contents as near as possible to end users, for instance, at E-BSs, equipped with data storage. In this way, the distance between two nodes decreases and the traffic requested by users is addressed to the nearest node. The communication is more faster. Surely, this is truth when the content the user is asking for is stored in the edge cache. If it is not, the request has to pass through the backhaul link to the higher level of servers and then the content can be delivered. In Figure 1.3, this architecture is presented: caches are located at distributed edge nodes , at **one hop from the end users** in order to minimize the network response time. Here, contents are stored according to a given placement algorithm.

Figure 1.3: Content Delivery Network with edge base stations

When a request occurs, the architecture avoids it has to pass through the network up to the core. Users can be earlier and better served. The better is the caching placement at E-BSs, the better users are served.

Nodes: Edge-BSs

The geographical servers' distribution around the world provides a caching infrastructure allowing high contents availability and high system performance.
Each edge server copies and caches contents of the core network.
A predefined amount of nodes has to be placed and connected strategically in the infrastructure and their locations have to be determined within a short distance to users [6], [10]. Caches with a limited memory are installed at E-BSs in order to have popular contents to clients.
When a user requires a stored content, it can be directly served. Otherwise, the request has to pass through the backhaul link (**cache miss**).
E-BSs are equipped with different interfaces, allowing them to interact with different communication systems. For example, in hybrid networks, which are going to be treated in Chapter 3, the interface for satellite communication is provided, in addition to the interface for radio access. The radio interface is related to a multihop unicast network in which the cached content has to go through multiple links and has to be transmitted individually. The satellite interface one is a multi/broadcast link which allows, in hybrid network, the satellite to fill the edge caches according to the placement algorithm. In the presented model, on Chapter 4, we consider caches equipped with two interfaces: one for mmWave and one for satellite.
Since edge caches have limited size, what is supposed to be stored is very important in terms of users' QoE. The performance optimization problem, in fact, passes through the definition of an optimal placement algorithm: the better placement algorithm, the better the system performance. Data storage is based on contents' popularity which is, in this case, the files' popularity computed by each edge-base station based on users requests. Different E-BSs can store very different contents since the popularity of each file can be very different in different portion of the network.
It this way the network is really users content oriented, being able to discriminate popular contents.

1.2 Caching techniques

The main focus of the caching algorithm research is to find the best approach to update the cache content. In other words, this area tries to answer these questions: which files to keep/remove and when to keep/remove files under different network architectures. Two different caching content approaches are available:

A **Off-line caching**
In off peak hours, the contents storage is done applying a given placement algorithm. It has to consider what was happening during the specific previous defined time and, if it is necessary, to update the cache content.

B **On-line caching**
In this approach, operations are performed on the fly and knowledge of future requests.
Any time a user requests a file, according to the adopted replacement algorithm, if the content has defined requisites to be stored, it is put directly into the cache such that, if another user asks for that content, it can be immediately served.
This approach is, clearly, more complicated but, if it is done in an intelligent manner, it achieves higher performance compared to the off-line caching. In fact, an on-line cache update is able to take in account variations in users' behaviour,creating a users content oriented network.

Since caches have limited size, not all contents can be stored. For this reason, **cache replacement policies** are needed and they play a key role. In fact, based on the replacement policy, contents are better or worst saved in caches, increasing or decreasing performance. Two approaches of cache content replacement policies are presented commonly in literature, based on single factor or multiple factors:

1. Single-factor replacement policies:

   - **FIFO**, fist-in-first-out;
   - **LRU**, Least Recently Used;
     The policy focuses on the least recently accessed content. Temporal locality is defined as type of principle of locality, which is, in computer science, a term for the phenomenon in which the same values are frequently accessed. It refers to the reuse of specific data, and/or resources, within a relatively small time duration. LRU influences temporal locality of reference-namely, that recently accessed objects are likely to be accessed again. It replaces the block in the cache that has not been used for the longest period of time. From the basics of temporal locality, the blocks that have been referenced in recent past will likely be referenced in the near future. This policy works well when there is a high temporal locality of references in the workload [11], [12].
     An extention of the classic LRU is treated in [13]; it is the Early Eviction LRU (EELRU); the algorithm evicts the blocks when it notes that

too many pages are being touched in a roughly cyclic pattern that is larger than the main memory.

- **LFU**, Least Frequently Used;
  The LFU policy replaces the least frequently used content in the cache [12]. If a page with high frequency will no longer be used, it will take up the cache and the useful pages have fewer buffer space. The LFU policy has several drawbacks: it pays almost no attention to recent history and does not adapt well to changing access patterns since it accumulates stale pages with high frequency counts that may no longer be useful [14, 15].

- **LFU-Aging**;
  This policy is an extension if the classic LFU. It is a frequency-based policy that tries to keep popular documents in the cache. When space is needed for a new file in the cache, LFU-Aging removes files with the lowest reference count. The policy is used periodically to reduce the reference counts of cached documents, so that formerly popular documents do not clutter the cache long after their popularity diminishes [16].

2. Multiple-factor replacement policies, like:

   - **GDS**, Greedy Dual Size;
     It combines temporal locality, size, and other cost information. The algorithm assigns a cost/size value to each cache block [17]. In the simplest case, the cost is set to 1 to maximize the hit but other costs such as latency, network bandwidth can be explored. GDS assigns a key value to each object. The key is computed as the objects reference count plus the cost information divided by its size. The algorithm takes into account recency for a block by inflating the key value (cost/size value) for an accessed block by the least value of currently cached blocks. The GDS-Aging version adds the cache age factor to the key factor. By adding the cache age factor, it limits the influence of previously popular documents.

   - **LRU-K**;
     It is a variant of the LRU policy considering the last access time and access frequency of a document. The last K access times of every document are recorded. When there is a necessity for cache document replacement, documents with access frequencies less than K will be replaced first. Subsequently, the policy replaces documents that have not

been requested recently, according to the LRU policy. If the time since the eviction of a document is greater than a defined metric, then the access records of the document are deleted [18], [17].

Moreover, there are the random replacement policies, RAND, choosing among all blocks in the cache with equal probability. It is the simple approach.

1.3 Popularity

Since not each file is equally accessed, the files popularity information is an important factor for effective cache replacement policies. It allows to understand which content is going to be most probably requested and how often. Therefore, based on the given placement algorithm, storing in the E-BSs the most popular contents achieves users' QoE improvement.
The local popularity is computed by each E-BS based on its view of the network and so based on its overseen users. Concerning users requests, some empirical observations can be done: local popularities tend to behave according to geographical positions and the considered time in which they are observed. It means users belonging to adjacent E-BSs at some point in the observing time require, more or less, the same contents, which are going to be very popular at that time in that region. Varying time as well as moving faraway from a certain E-BS, the files' popularity distribution changes with high probability. The popularity modelling has been largely investigated, since it is so important for system performance improvements. Focusing on Video contents, it is lage largely accepted in literature [19], their popularity follows Zipf distribution. 10% of the online videos account for nearly 80% of the views, while the remaining 90% of the video account for only a total 20% of views. Content popularity estimation, selection, and delivery are studied in [20, 21, 22]. The authors in [22] consider a network comprised of a macrocell base station and multiple small cell base stations where each of them has a limited cache capacity and serves a group of users. The authors minimize the content delivery delay to the users. To do this, first, a training period is carried out to put the users with similar content request into a group and each group is assigned to a small cell base station. After clustering the users, learning coarse algorithm is used at each small cell base station to locally learn the content popularity and update the cache accordingly. The authors of [20] consider multiple small cell base stations where each of them is connected to the core network using a limited capacity backhaul link. The authors use the social device-to-device interactions background of each user to improve the estimation of the file popularity matrix since the user ratings

are not enough in the small base stations and this leads to a highly sparse popularity matrix. This process is refereed to as the transfer learning. The estimated matrix is used for content placement in the off-peak hours to minimize the backhaul traffic by considering the storage limit, estimated popularity, and backhaul capacity. The proposed transfer learning approach is compared to the ground truth algorithm, uses the perfect popularity matrix to store the most popular content, random caching algorithm, contents are cached in a uniform random fashion, and collaborative learning algorithm. It is shown that the transfer learning approach outperforms the benchmarks in terms of backhaul off loading and user experience. As a practical approach, [21] re-examines [20] by using statistical machine learning tools to estimate the file popularity using enormous users' data from a mobile company. The work minimizes the average backhaul load by considering the file size, file bit rate, backhaul link capacity, wireless link capacity, cached content, and user quality of experience. The authors run big data analysis on users' data and show it follows a Zipf-like distribution [17]. The big data approach is compared to the collaborative filtering method and is shown to have a better performance in terms of backhaul offloading and improving user experiences.

Advantages and problems of CDNs, edge

CDNs with edge nodes have brought high advantages to the traffic management, when a cache miss doesn't occur. Caching contents at edge base station can, in fact, improve the delivery phase, when a content is stored. Most important CDN's **benefits** are [23]:

- Transparency;
  Users don't realize the structure has been changed. They send requests and receive services but they don't know through which links requests pass. The services' quality is not compromised.

- Latency reduction perceived by users;
  The contents' delivery is faster since they are stored in geographical nearest edge servers. Users ask for contents: if they are stored in the local cache, the request can be directly delivered, avoiding to pass through longer links.

- Increasing redundancy and scalability;
  The first one is guaranteed by storing the same content in various edge nodes. The second one is related to the possibility of adding new server or removing old ones if it is necessary, without affecting system performance.

However, the caching policy is based on local popularity distributions. This means each E-BS evaluates a content as popular or not only based on its local view of the network. This doesn't take in account popularity is related not only to local factors but also to global trends. These cannot be evaluated by the E-BS. To evaluate global popularity, an higher layer in the network is needed.
Hybrid networks, based on the cooperation between standard CDN and satellite overlay can solve this problem. In fact, thanks to the satellite, a global view of the network is available, creating a users content oriented network in the global meaning.

In literature, a lot of works presents studies and implementation proposals of these networks. The application of satellite communications due to high bandwidth and wide are coverage is investigated in feeding several network caches using, at the same time, broad/multicast [24], [1]. The work of [24] proposes using the broad/multi-cast ability of the satellite to send the requested contents directely to end users. The authors of [1] consider a network comprised of proxy servers with individual cache storages. The proxy servers and the satellite gateway are connected to the content delivery network. If the requested content is not available in the cache, the request is routed from the satellite to base station. Based on the number of requests, the file size, and the available satellite bandwidth, the request will be admitted or dropped. The satellite multicasts the file global popularity as well as the requested contents. Each server uses the local and global file popularity to update the cache. Through two case studies, the authors show that the satellite multicasting can be effective in terrestrial backhaul off loading. Also, this technique is scalable in terms of the the cache storage and the coverage area of the content deliver network.

Chapter 2

Hybrid networks

An hybrid satellite/terrestrial system is a system employing interconnected satellite and terrestrial components operating independently of each other. In such systems the satellite and terrestrial components have separate network management systems and do not necessarily operate in the same frequency bands.
Recently, terrestrial and satellite networks have been fighting each others to establish which one is the best at, for examples, broadcasting television and backhauling data in remote areas, mobile telephony and aircraft telecommunications services. In this competition, satellite networks have been suffering more than others because of, the high costs, with respect to the benefits that could have been achieved. Current CDNs store contents at the edge of the network according to files' popularities, not considering popular trends. The caches are filled using unicast transmissions, which means significant time is needed for the placement of the contents and high network resources allocations.Therefore the persistent growth in data requirements of terrestrial networks has forced to think about alternative solutions in order to increase the quality of offered services to clients. A promising way to feed the cached with popular contents is to use a satellite overlay, which can reach a very large number of receivers with low cost, can use its global view of the network to know global trend in users behaviours [25]. The satellite hybrid architecture aims to jointly benefit from the advantages and capabilities of both terrestrial and satellite telecommunication systems, supporting a diversity of services, and especially efficient multicast and broadcast services [26]. On the other hand, building an hybrid network brings along some issues as the management of point-to-point, point-to-multipoint and multipoint-to-point communications, the routing traffic and the evaluation of different criteria like cost, quality of services, parameter definitions and application type [27]. European Telecommunication Standards Institute (ETSI) works are focused on hybrid satellite systems, which are treated in

different projects such the **SatNex, "Satellite Communications Network of Excellence"**, who is also founding this work.

2.1 Motivations

Hybrid networks are an efficient and cost-effective solution to employ satellite communications not only for broadcast and multicast services but also for mobile services. In fact, the recent growth in data demands along with the increasing number of users accessing to the network, make terrestrial networks suffering and so regarding to the decreasing performance. Nowadays, wherever a user requires a content to the network, the CDN approach is used. This approach contents are stored at the edge of the network and delivered to the user when it is needed. Problems occur when there is a cache miss. In this case, the user request has to pass through the backhaul link, which is the most affected by increasing data traffic. Avoiding cache miss can be achieved storing contents, at edge caches, appropriately. One way is to save contents according to their popularity by to the edge base stations (E-BSs). Howerver, E-BSs has only a practical knowledge of the network and thus their local view can be much different from the global one. Indeed, in this way the placement doesn't consider the global popularity distribution of files that can influence local users behaviours. Moreover, contents are sent from the server to the E-BSs via unicast, making backhaul links very busy. The main idea of hybrid network is to jiontly distribute the load among different networks, achieving improvements in users' quality of experience (QoE), due to the balance interoperability of the networks. This can be done splitting data requests among different parts of the network as well as among different networks avoiding a congestion [28]. The hybrid approach focuses on the interoperability of satellite and terrestrial components which facilitates long radio link ranges and countrywide/continent-wide service coverage in a cost efficient manner. In [26], different level of integration between the networks are studied as necessary for the cooperation between satellite broadband networks and terrestrial wireless access networks (e.g. WiFi, WiMAX, 3G, LTE). The management of the cooperation among different wireless communication systems is treated in [29], introducing satellite to help pre-existing terrestrial network, improving users' QoE . Caching has emerged as a promising way to reduce client-perceived latency and network resource requirements and there are plenty of works about it. In [25] an analytical model to study the performance of a cache-satellite network is developed, which do not require any inter-cache cooperating protocol, achieving significantly reduc-

tions on the cost of filling caches. The integration of the satellite component into the 5G ecosystem is investigated in [30] focusing on mobile backhauling, where satellite capacity is used to support the terrestrial backhauling infrastructure, not only in rural areas, but also for making traffic delivery to radio access network (RAN) nodes more efficient. An analysis of the capacity and traffic management strategies in hybrid satellite-terrestrial mobile backhauling networks that rely on Software Defined Networking (SDN) is proposed. In [31] and [32] the cooperation between satellite and Transfer Control Protocol (TCP) is investigated. The hybrid satellite and terrestrial network based on the soft defined architecture is proposed from a perspective of 5G in [33] investigating the behaviour an end-to-end architecture. The performance are analysed based on the stochastic geometry, where the satellite and the relay cooperatively transmit signals using Alamouti space-time coding scheme. In [34] an hybrid system in which satellite cooperate with terrestrial CDNs is presented. [29] are one of the most crucial points in spectrum utilization: a multifrequency network (MFN), or a single frequency network (SFN). The first enables two indipendent radio access systems at different frequency bands, allowing two communication components not interfering with each other. The second enables frequency sharing leading to a more efficient spectrum utilization at a cost of increase of the influence. The handover functionality is investigated in [26], since it becomes more complicated in hybrid networks because also inter-system handover shall be performed. The handover initiation is more tricky and depends in addition upon load balancing in the system, or can even depend on more complicated procedures such as different cost functions, network state and specific connection admission control procedures with forced handover. The protocol convergence shall be achieved in order to facilitate these handover procedures and the design of multi-mode terminals [26]. For the issues of authentication, security and billing shall be carefully and consistently addressed. End-to-end security shall be provided in a coherent way and at different network layers. The amount of traffic carried over different sections of the hybrid system shall also be carefully considered, since prices may largely vary from terrestrial to satellite systems [29]. To summarize a bit, implementing an hybrid network brings a lot of **advantages** that are listed down hereafter:

- service coverage extension; This is due to the large satellite footprint that can reach users even located in remote positions.
- broader range of service provisioning and lower costs for customers and operators; The cooperation between systems allow users taking the best from both.

- rapid and infrastructure independent service deployment; The satellite can oversee the great part of users, apart their distribution and location.
- increase of the QoE offered by operators; Users are better and early served. This increase their QoE as well as system performance.
- optimal usage of the resources of the telecommunications networks. This happens by selecting the most appropriate network for the transmission of each service [1], and by decreasing the traffic load in congested terrestrial areas with the transfer of part of the traffic over the satellite systems with appropriate load balancing mechanisms;
- increase in service availability and resilience; More services are available since the networks is not overloaded: users asking for contents are served early.
- energy efficient of the operational/overall investment cost when deploying a new service;
- Optimal energy architecture; It is done by taking advantage of the broadcast/multicast capability of the solar powered satellite infrastructure;

2.2 Satellite overlay: benefits and problems

The increasing demand on high-speed streaming applications, spread all over the world. Geostationary (GEO) satellite systems considerates as potential means of communications [35]. Compared to the terrestrial networks, satellite networks offer significant advantages in terms of cognitive capabilities, which maximize the utilization of radio resources, the larger spatial coverage, the ability to offload and cache content and realize more efficient multicast delivery [26]. Based on the soft defined features, splitting requests in hybrid satellite and terrestrial network could be one of the key to enable next generation systems to create various customized scheduling and allocation schemes while maintaining coverage. Hybrid networks can, indeed, provide end user devices with adapted and scalable capacity, network coverage and access and satisfy various quality-of-service constraints [36]. The choice of a satellite telecommunications system rather than a terrestrial one is usually driven by satellite natural capabilities and advantages [37],:

- *high bandwidth.*

[1] multicast or broadcast services are preferentially carried over satellite systems, whereas conventional or point to point services are transferred over the terrestrial networks

- *very large coverage areas.*
- *inherent multicasting and broadcasting capabilities.* Satellites can send broadcast or multicast contents, based on what the system needs.
- *cost effectiveness.* Cost of satellite capacity does not increase with the number of users/receive sites, or with the distance between peers. Whether crossing continents or staying local, satellite connection cost is distance insensitive.
- *global availability.* Communications satellites cover all land masses and this is the reason way there is growing market serving maritime and aeronautical applications. Customers in rural and remote regions all around the world who cannot obtain high speed Internet access from a terrestrial provider are increasingly relying on satellite services.
- *superior reliability.*Satellite communications are standalone, i.e. they can operate independently from terrestrial infrastructure. When terrestrial outages occur from man-made or natural events, satellite connections keeps working.
- *robustness.*
- *immediacy and scalability.* Additional receive sites, or network nodes, can readily be added, sometimes within a few hours. All it needed is a ground-based equipment. Satellite has proven its value as a provider of "instant infrastructure" for commercial, government and emergency relief communications.
- *versatility.* Satellites effectively support on a global basis all forms of communications ranging from simple point-of-sale validation to bandwidth intensive multimedia applications. Satellite solutions are highly flexible and can operate independently or as part of a larger network.

For all these reasons, satellite seems to be the most appropriate system to serve different areas like coverage in planes, navy ships, hostile environments and so on. As any communication system, also satellite presents some disadvantages:

- *Technological complexity.*
- *High costs.* It is related to the huge initial cost. In addition to the cost of building one of satellite devices, there is also the cost of launching the satellite into space.
- *Repair of satellite is almost impossible.*

- *Delay.* It is the length of time it takes for the satellite to communicate with Earth. This delay can vary greatly. More than anything else, this is caused by the huge distance over which the satellite must send the signal.

To overcome the delays over reverse links, several works have proposed different solutions [38], [39]. A network coded Automatic Repeat-reQuest (ARQ) protocol for broadcast streaming applications over hybrid satellite systems is proposed in [35]. It exploits the abilities of full deterministic network coding in generating efficient proactive retransmission packets. These packets are transmitted right after their original packets without prior knowledge of their loss status. This greatly improves the average packet delay performance in such high Round Trip Time (RTT) systems.

2.3 Hybrid scenarios

Different network scenarios can be created implementing hybrid architectures, based on a satellite architecture. Among them it is worth to mention:

1. *Broadcast services*
   Satellite and a terrestrial components must be combined to broadcast media contents, like radio and TV programs, to the end users. Broadcast services generally requires an unidirectional satellite link between the network component and the satellite segment, mainly because the high bandwidth consumption on the downlink (from network to user equipment) whereas the bandwidth requirement on the uplink is much less and hence, can be managed by the terrestrial component of the hybrid architecture. It is the case of Video on Demand requests in which high bandwidth demanding services is served through satellite. Each end user terminal should be able to receive both satellite and terrestrial signals for smooth service continuity over the coverage and to allow combining terrestrial and satellite signals when both have acceptable quality in order to achieve diversity gains.

2. *Telecom access network services*
   Satellite systems can advantageously be used as additional Satellite Radio Access Network (S-RAN). The S-RAN is a collaborative extension of the classical terrestrial cellular 3G RAN, which allows to extend the coverage of classical cellular wireless RANs which are not covering remote areas. Load balancing and traffic differentiation can be performed, taking into account quality of service criteria as well as the application type (N point-to-M point

links) while deciding which communication segment, satellite or terrestrial, to be employed. The satellite network operates in both forward and return directions to provide an alternative access network for a mobile user equipment. The same equipment can also latch to a terrestrial cellular network to access the same or other services. The spectrum used by the satellite network in the hybrid architecture is a frequency band allocated to Mobile Satellite Services (MSS) in L-band (1.5 GHz and 1.6 GHz) or S-band (1.9 GHz and 2 GHz). Using those bands, the same user equipment can work without any interruption of services in the satellite as well as in the terrestrial segment of the hybrid architecture, optimizing the resources utilization, that allows the provider to employ the terrestrial infrastructure for services/applications that require less bandwidth. For specific interactive Multicast/Broadcast services, the satellite can be preferred for the forward link while the terrestrial component is used for handling service requests. Similar implementation can be done for the fixed terminal equipments where a bidirectional satellite link can be employed in a collaborative manner with the existing terrestrial network, building a so-called ADSL broadband network.

3. *Backbone technology*
   In remote areas or hostile environments like maritime, military operations or reconnaissance missions where it is very difficult to setup a terrestrial network hybrid network could be used as a backhaul to connect two or more terrestrial points of the network. This can also be employed in case of emergency group of users involved in rescue operations in remote areas.

4. *Content delivery services*
   As already discussed in Chapter 2, a CDN aims at distributing multimedia contents towards data centers in the transport network so that the content is served to end users with higher service availability and performance. This allows to avoid a website to become virtually unreachable because too many people are hitting it or reducing the general load on websites servers in general. It also lessens the demands on the network backbone and to reduce infrastructure investments. Nowadays, CDNs serve a large fraction of the Internet content, including web objects (text, graphics, URLs and scripts), downloadable objects (media files, software, documents), applications (e-commerce, portals), live streaming, on demand streaming, and social networks. With the increase in the number of data centers, the inherent cost effective multicast/broadcast capability of satellite becomes more relevant for CDNs. Satellite systems can be used efficiently in CDN networks to feed

CDN servers and caches them thanks to multicasting. The benefits of using satellites include the transport of high volumes of bulk data, between any CDN nodes within the satellite coverage and also offload terrestrial networks so that they can handle more easily short haul connections, thus requiring small delays (time-sensitive services).
The terrestrial component of the hybrid architecture deployed for CDNs is typically made of wireline/wireless transport/ networks with data centers to distribute the content directly to fixed and mobile end users. The satellite component is usually based on GEO stationary satellites and connects the data centers to the service platforme citechatzinotas2015cooperative.

Hybrid network can be used in a huge variety of applications. The most important are peer-to-peer, near Video on Demand, multicast application in 3G cellular systems, broadcast streaming in which a reliable delivery of packets is required to occur at all receivers before a certain deadline beyond which these packets become useless [35], integrated DVB-SH systems , hybrid TerreStar system, cognitive hybrid communications systems (coexistence of terrestrial and satellite systems, satellite assisted terrestrial network, combined satellite-terrestrial system with cognitive radio techniques).

2.4 Hybrid system performance and challenges

Terrestrial solution may be good for fixed receivers where an established wired connection can be employed. However, it may not be quite adequate over terrestrial wireless networks for mobile receivers as it will greatly overload these networks with retransmission packets especially in broadcast scenarios with large number of receivers, employing low bit rate terrestrial reverse links only for packet acknowledgments [35].
In a hybrid network, composed of a satelliteand a terrestrial segment, a lot of improvements can be achieved with respect to the case conseidering both segment disjointely. presented [29]. To name a few:

- supporting diversity of services (especially efficient multicast and broadcast);
- enabling users to be anywhere at any time connected by ubiquitously extending the coverage area of the telecommunication systems, at the best price and with the best possible connection [35];
- optimizing resources utilization by selecting the most appropriate network for the transmission of each service and by decreasing the traffic load in congested

terrestrial areas offloading part of the traffic over the satellite systems thanks to appropriate load balancing mechanisms;

- throughput, power consumption, spectral and energy efficiency, and coverage probability improvements;
- efficient resource management mechanism to achieve the orchestration of the network resources according to the context and requirement of services based on the network deployment for future 5G wireless network.

2.5 Satellite overlay over CDNs

There are different ways to combine satellite and terrestrial networks, each one leading to a particular hybrid system. In these hybrid systems, some specific network functionalities should be supported focusing, for example, on load balancing and QoS support as well as on handover. It is important to design a system architecture that could be easily adapted to different networks configurations and to properly manage the network deployment phases[40]. A simple architecture implementation of an hybrid satellite/terrestrial network is presented in Figure 2.1:

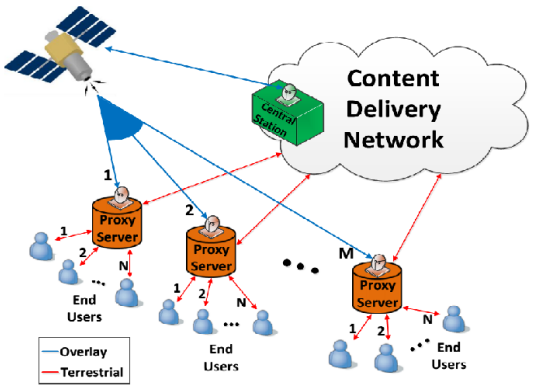

Figure 2.1: CDN topology with satellite overlay [1]

This architecture was first proposed in [1] which aim at leveraging network scalability in presence of increasing data demand. n the presented architecture, the main components are:

- a space segment which includes the satellites, commonly a GEO satellite;
- a terrestrial segment which includes stations connected through the Internet, providing the link between the satellite system and satellite terminal segment; In general the terrestrial part is made by multi-hop wireless network with self-healing and self-configuring capabilities, dynamically self-organized, with the nodes of the network automatically establishing an ad-hoc network and maintaining the mesh connectivity [40]. It is very important to harmonize the core network structures for accomplishing seamless cooperation between the terrestrial and the satellite segments.
- a terminal segment which includes a terrestrial terminal segment, composed of user terminals such as satellite phones that provide direct satellite access to end users, and a terrestrial terminal segment that is composed of end user terminals.

The system scalability is investigated, creating a satellite-based overlay for existing terrestrial CDNs. The potential benefits of multicasting communication via satellite is evaluated through simulations on two case studies - Cellular and Video on Demand -. Results show that multicasting has the potential to provide significant bandwidth reductions from terrestrial-based unicast solutions. In addition, it is scalable in terms of both cache storage and coverage area of the CDN. There are a lot of future challenges related to these hybrid networks, mainly due to the expected development in satellites, to improve network performance and users' QoE. They are related to the long term-distance links in satellite systems, leading to long transmission delays compared with terrestrial communications [29]. Moreover, since satellites are designed to operate for a long period, technological solutions are needed to be defined before beginning of services.

Chapter 3

System Model

The most important aim in the proposed model is, conceptually, the cooperation of two different networks, focusing on both satellite and terrestrial network users' views and therefore on the files' popularity which is based on users' requests and can be estimated by satellite and terrestrial networks. We consider a system model in which a satellite overlay covers a pre-existing heterogeneous terrestrial network for content delivery implementing a satellite-assisted caching to provide improved cache-feeding performance and user experience. We focus on offline satellite-assisted caching where a monobeam satellite first and a multibeam satellite later, along with the terrestrial network, are used to feed the caches of the E-BSs, which are located at the edge of the network, at one hop from the end users. The edge caches are equipped with two interfaces, one for mmWave terrestrial backhauling (a multihop unicast network) and the other for satellite backhauling (broadcast or multicast link).

To understand which content is adapt to be stored and where it can take place, contents' popularity information is needed, both local and global. Using its local coverage, E-BS learns from its users requests and computes the local popularity distribution.

On its hand, using its high coverage and high bandwidth, the satellite has a global look on users' requests since it knows the local popularity computed by E-BS and, averaging all of them, it can estimate the global popularity distribution. Once both popularity types are available, we consider the placement phase in which satellite broadcast and the terrestrial unicast transmission links are both involved to feed the caches. To evaluate system performance, we focus on hit ratio, which the number of stored requests, and placement time, which the needed time to fill the caches. In order to minimize the placement time and maximize the hit ratio, the important question is how to decide which files to put in the caches and which

one to broadcast or to unicast. Fixing these objectives and assuming the same number of users as well as the same cache size per E-BS, we implemented three different placement approaches, in order to verify the hybrid placement implements intelligent caching, achieving performance improvements:

A. **Global Placement**
   Considering the global popularity distribution, the caches are filled with the most popular contents. This distribution is available since the satellite can have a whole view of the network and so it knows the local popularities calculated by each E-BS. It averages them, achieving the global trend. This approach doesn't provide significant advantages in terms of hit ratio because stored contents can be very different from local requests in some E-BS. In this case, each request must pass through the backhaul link. In terms of placement time, instead, the algorithm is very convenient thanks to satellite characteristics: it can simultaneously send to all E-BSs higher global popular contents supposed.

B. **Local Placement**
   Only the first files with the highest local popularity values are stored in caches. Each E-BS can calculate files popularity basing on its users' requests. Since requests are locally determined, based on local users behaviours, and caches are filled with contents having highest probability to be required, the algorithm provides the best performance in terms of hit ratio. On the other hands, in terms of placement time, performance is not so good because, to fill caches, each content has to pass through the terrestrial unicast links.

C. **Hybrid Placement**
   We focus on creating an efficient algorithm in which the network can take advantages from both systems. This is possible when caching is done considering local and global popularities, since users local behaviours are not decorrelated from the global trend. The proposed hybrid approach, moreover, combines satellite multicast/broadcast with terrestrial unicast transmissions to fill caches, sending, respectively, global and local popular contents. In Section 3.4 the implemented algorithm will be explained in detail. There, the only-global and the only-local placements are used as baseline solution while the hybrid investigates two methodologies, one centralized and one decentralized.

The presented can be developed in terms of users' distribution, users behaviour prediction and data traffic modelling as well as optimal caches' positioning. Then,

the so-proposed hybrid architecture could represent an attractive challenge for future works and improvements. All the evaluations are considered based on the same number and distribution of users at each of E-BSs, the same satellite's bandwidth, the same local popularity distributions and the same placement algorithm. In a schematic representation, our model is made by a satellite overlay over a CDN which cooperate to fill edge caches in order to better serve the users. Hereafter each component will be to be briefly illustrated.

3.1 Scenario

In our model, pre-existing CDNs are the terrestrial network level. As explained in Chapter 2., in these networks, edge caches at E-BSs store contents according to a given placement algorithm. In this thesis, two different scenarios have been considered: the first one, shown in Figure 3.1, represents the monobeam satellite overlay and the second one, shown in Figure 3.2, represents the multibeam satellite overlay. Each of them concerns a different approach in calculating the files' global popularity distribution.
The monobeam satellite uses its entire view of the network, learning from each E-BS while the multibeam one uses its global view of the sub-network in each beam, since only a portion of the entire network is illuminated by the satellite stack. In this latter case, the global view is intended as the whole view of the sub-network in the local section.

3.1.1 Satellite Overlay

In the hybrid architecture, the satellite produces benefits because of its wide area coverage, broad bandwidth and its multicasting features which depends on the satellite type we refer to. In fact, based on this, the employed satellite can broadcast contents to all E-BSs or multicast to groups of them. A monobeam, first, and a multibeam, later, satellite are considered: they differ from global popularity calculation.

Monobeam Satellite

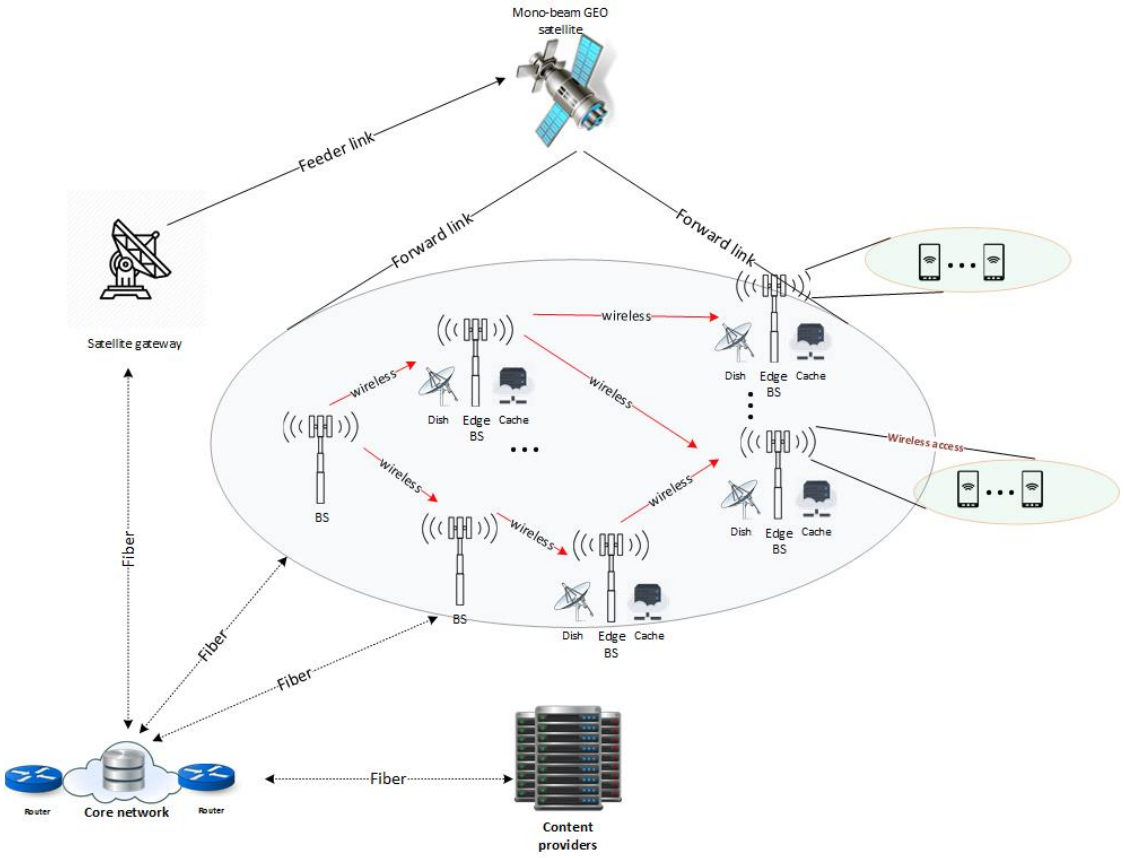

Figure 3.1: Monobeam satellite over an heterogeneous CDN

Using the monobeam satellite, Figure 3.1, the files global popularity distribution is computed averaging the local popularities. The satellite learns them from all the E-BSs therefore it broadcasts contents to E-BSs, according to the caching algorithm used. If it is decentralized, which means at each edge cache different amount of global contents are stored (in each E-BS a local threshold is used), it sends via broadcast contents at each E-BS and they evaluate which one to memorize and which one not. Otherwise, when the algorithm is centralized, at each E-BS the same amount of global contents is sent via broadcast to all E-BSs and stored by them. This is very convenient in terms of placement time and resource occupation.

Multibeam Satellite

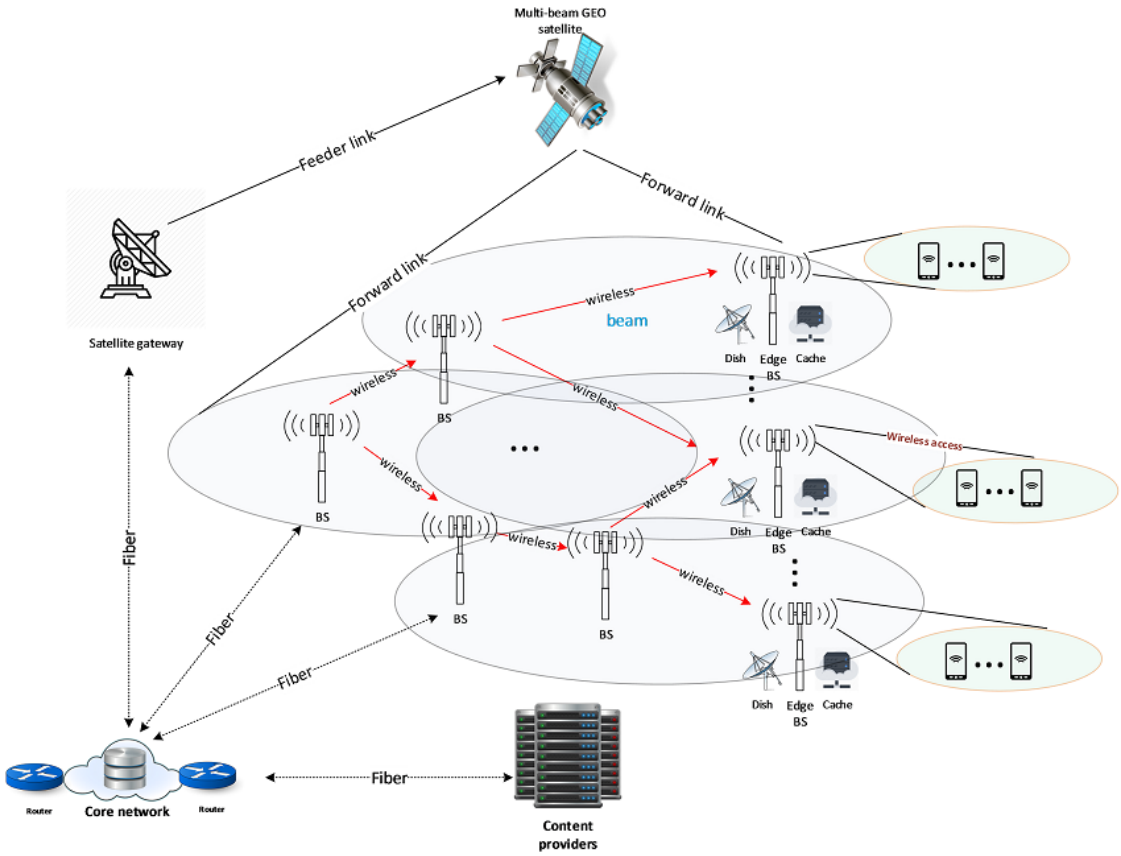

Figure 3.2: Multibeam satellite over an heterogeneous CDN

The multibeam scenario is characterized by disjointed lighted up network portions by the satellite, the beams indeed. Each sheaf can be interpreted like a global part of the network, locally. In each beam in fact, the satellite learns local popularities from E-BSs belonging to that beam and calculates the files' global popularity distribution averaging over them. This means, basically, a new level of popularity is introduced. Also, in this case, once we have the global popularity and the algorithm placement, edge caches can be filled according to the algorithm, using unicast or multicast transmissions.

3.1.2 Users

Users are supposed to be uniformly distributed. Each BS therefore serves the same number of users. Users behaviour modelling represents one of the future challenges in hybrid satellite network. In fact, once we know their preferences we know which kind of files they are going to request and, consequently, save them near by end users, improving the delivery. This is very complicated to model,

mainly in offline caching, in which the replacement is done not in short time, like in the online approach, not using changes in users behaviours during time. Recent studies show that users in the same geographical area, most likely, are interested in same kind of contents. Starting from this, it is less difficult to define a files' popularity distribution, which depends on the geographic place we are observing as well as other more general trends, such as some contents becoming very famous all over the world. Exactly, users behaviours mainly depend on local factors but global trends on file popularity influence them: a file very popular globally is likely to be very popular also locally, even not everywhere. For that reason in this work we focus on files' popularity distribution in terms of modelling.

3.2 Files' popularity

The files' popularity is very important in order to understand which content is more requested then others in order to better fill caches and consequentially to serve users in the best manner. In the presented model, it is supposed files popularity distribution follows the so called **Zipf-like law**, described afterwords.

3.2.1 The Zipf-like law

The files popularity modelling has been largely treated in literature all over the years [1], [12], [17], [34]. Its trend has been defined, basically, based on empirical observations and, nowadays, it is accepted quite commonly that the popularity follows the Zipf-like law, which is a more general Zipf law's modification. The common Zipf-law, once events are ranked with respect to the frequency of presentation, provides the tie between the event's frequency of presentation and its rank. It is a general law applicable to many systems such as the relationship between words in a text and their frequency of use. According to the Zipf law, **the files popularity distribution is obtained starting from the presentation probability of each file,** $p(i)$ **and its rank** i:

$$p(i) = \frac{1}{i}. \tag{3.1}$$

The relative probability of a file is inversely proportional to the rank assigned to that file: the most frequent request file has rank 1. Basically, the rank is related to the frequency by which the file is accessed.
In the **Zipf-like distribution**, instead, the relative probability of a request for

the $i-th$ most popular file is so defined:

$$p(i) = \frac{1}{i^{\alpha}} \tag{3.2}$$

with α typically taking on some value less than unity [17]. Very often it is chosen in the range [0.5, 1]. Since the (3.2) doesn't give a probability distribution, a normalization is needed in order to obtain:

$$\sum_{k=1}^{N} p(i) = 1 \tag{3.3}$$

Letting $\Omega = \frac{1}{\sum_{k=1}^{N} p(i)}$, the **Zipf-like law** can be expressed:

$$p(i) = \frac{\boldsymbol{\Omega}}{i^{\boldsymbol{\alpha}}} \tag{3.4}$$

Global and local distributions follow this law and their trend is shown in Figure 3.3. Each file has, in the two different distributions, different files rank: in the local case, even if the shape is the same, the most popular file's rank could be very different from one.

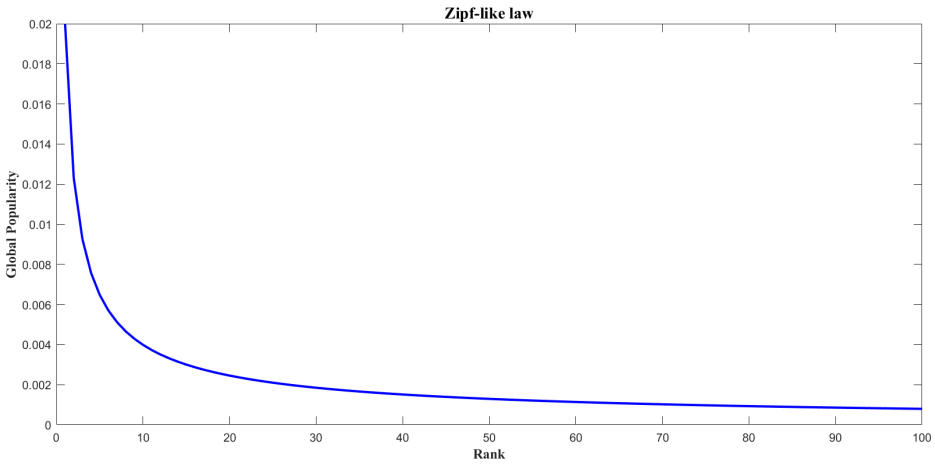

Figure 3.3: Zipf-like distribution

3.3 Centralized and decentralized caching

Two different caching approaches are studied, for the network point of view: the centralized and the decentralized caching.
In the first one, an unique threshold is fixed per each edge cache. These are so filled with the same amount of globally popular files and with the same amount of

locally evaluated contents. The satellite sends a number of contents to store which is the same in each cache all over the network; edge caches contents will be different among them because of local contents, having different popularities compared to the global ones and different local popularities at each cache.
The decentralized approach, instead, consists in evaluating, base station by base station, each local popularity distribution value. Averaging global popularity values, an average value is obtained. Only contents with popularity values higher than the average one are stored, as long as there is space in the memory. If there is no enough space, only files with first higher values will be stored, since the cache is filled. If number of files with higher popularity values is not enough to fill the cache, this will be filled with files having first higher global popularity values, if not already stored, until the cache doesn't have available space any more. In a so presented manner, caches are proactive: each one decides what to do, based on the placement algorithm and on its users requests. Obviously this gives better results in terms of hit ratio.

3.4 Placement algorithm

Three different placement algorithms are implemented in order to compare results in terms of hit ratio and placement time: the local, the global and the hybrid algorithms. First two are used like as baselines: while the local placement gives the best performance in terms of hit ratio, it is the worst in terms of placement time, since unicast transmissions are employed. The global placement is exactly the contrary: it is the best for the placement time and the worst for the hit ratio. The hybrid placement algorithm is the new approach in which the satellite overlay collaborates with terrestrial network to fill the caches, improving QoE. Employing hybrid placement, if in one side the hit ratio suffers a lightweight decrease, using the satellite achieves reductions in terms of placement time. The algorithms take place when traffic requirements are scars and caching update can be done, within all network, avoiding overloading (offline caching). A so presented solution doesn't work in real time applications. Based on what is supposed to be offered to the clients, the optimal working point is identifiable as the reasonable trade off between hit ratio and placement time. In Figure 3.4, it is shown how the three algorithms fill the cache. There is also represented an optimal hybrid placement, in which the cache contains more locally popular contents than globally popular ones: this, as explained in Chapter 4, increases the hit ratio.

3.4.1 Hybrid placement

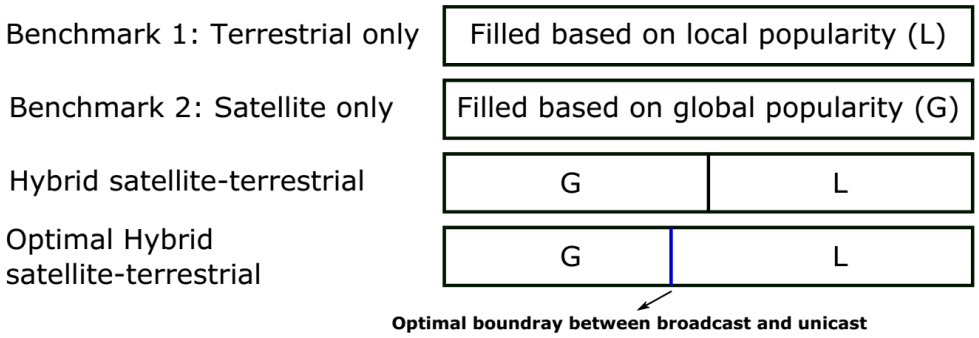

Figure 3.4: Schematic of the benchmark and proposed caching schemes

The hybrid placement algorithm is the main part of the developed work in order to verify and show satellite actually contributes to increase perceived quality services at final users.
To do this, both unicast and multi/broadcast capabilities have to be employed, distinguishing which contents are satellite expertise and which are the terrestrial network one.

Two different approaches are investigated: the centralized and the decentralized ones.

- Centralized placement
  Assuming each cache has the same size, the threshold (λ) has been defined as $\frac{CacheSize}{2}$. First λ stored contents are the first λ files with the highest global popularities. These are sent to caches by satellite, via broadcast or multicast according to the involved satellite.
  Second λ in-cache contents are the first λ files with the highest local popularities, if not already insert in the first part by the satellite. On the contrary, caches will be filled with higher local popular contents not stored yet, since there is enough space. Such contents are sent to caches using the terrestrial link, via unicast, cache by cache.
  The obtained results, using hybrid placement, allow interesting observations: the hit ratio trend has higher values than the global-only approach but less than the local-only one. The placement time is wors than the global time needed but it is better than the local approach.
  This means that using correctly the two systems, without any changes in the architecture or implementation additional costs, an optimal working point, in

terms of system performance, can be found. To do this, a trade-off between the two metrics, hit ratio and placement time, has to be fixed, according to the application requirements. The better the caching is done, the better the results are. Thus, the optimal placement algorithm passes through the investigation of an optimal threshold, allowing to obtain requested performance.

- Decentralized placement
  The second investigated hybrid placement algorithm represents a decentralized approach in which decisions regarding contents to store is done cache by cache. An average popularity value,G_{av}, is calculated averaging global popularity values and its used to define the threshold, th, in each cache.
  Fixing G_{av}, at each cache, the algorithm stores files having global popularity values higher than G_{av}. The remaining memory is filled using local popularity, taking contents having values higher than G_{av}, if not already stored. If there is still space, caches are filled using local popularity values until the cache is full. This achieves best performance in terms on hit ratio but it is not so advantageous in terms of placement time.

Chapter 4

Algorithm implementation and simulations

In order to show if hybrid satellite-assisted caching network is better than others caching techniques, simulations are created with the computing environment MATLAB.
In the first step, knowing the files' popularity follows the Zipf-like law, a reference distribution is generated, based on the theoretical popularity distribution, with respect to files rank. Starting from the reference distribution, the local popularities are achievable applying a semi-random permutation of reference ranks. Therefore the real global popularity can be computed averaging the local popularities. In the developed code, the placement algorithms are implemented and applied to fill caches. Local users requests are generated, based on local popularity distribution, in a semi-random way. The hit ratio and the placement time, obtained per each implemented placement, are shown in graphs. This allows the comparison among algorithms performance in order to decide which is the best, based on the required QoE. Obviously, this approach is valid both for monobeam and for multibeam satellite. In the latest case the global popularity is calculated averaging local popularities available at each E-BS belonging to that beam. The global view is interpreted, in the multibem scanario, as the global view of a sub-network.
The implemented algorithm can be summarized as follows:

Data:
N, number of files
K, number of caches
M, number of beams
C, Cache Size
Permutation Length
α, Zipf-like law parameter
while $t <=$ *updating time* **do**
    Reference Distribution Generation;
    Local Popularity Algorithm;
    Global Distribution Calculation;
    Placement Algorithms (Local,Global,Hybrid) ;
end
Users' Requests Generation;
Hit Ratio Calculation;
Time Placement Calculation;
Result:
Local Popularity Distribution
Global Popularity Distribution
In-Cache Stored Contents
Hit Ratio
Time Placement

Algorithm 1: Algorithm Synthesis

Both presented scenarios have been implemented using following parameters:

PARAMETER	VALUE
N, number of files in the library	10^4
File size	30 M byte
K, number of caches (E-BSs)	8000
α, Zipf-law parameter	[0 1]
Cache Size	[N/10 N]
NoR, number of users requests	1000
M, number of Beams	100
X, number of caches in each beam	K/M
permutation length	$N/2$
Threshold	$C/2$
Satellite throughput	50 M byte/s
Terrestrial throughput	50 M byte/s

Table 4.1: Simulations parameters

4.1 Reference popularity distribution

Supposing there are N available files in the library, sorted according to the files' rank from the most popular (rank 1) to the less popular (rank N), we generate a reference distribution, applying the Zipf-like law to the files' rank. Exactly, in our simulations, at the beginning the files' rank coincides with the files' index, which refers to the position of the file in the vector. The Zipf-like law is applied to the rank vector, using an α parameter belonging, quite commonly, in the range [0.5, 1]. The **reference distribution** is in this way obtained and it is represented in Figure: 4.1, in a logarithmic scale.

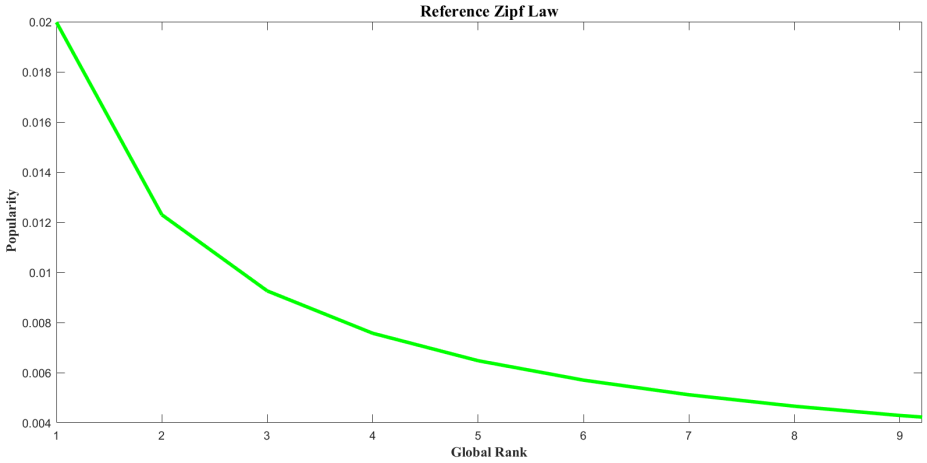

Figure 4.1: Reference Distribution

Different α parameters do not influence the file rank but only the file popularity values. If α is near to 1, the popularity values are more high, as it is shown in Figure 4.2, meaning most popular files have very high values of popularity, since they have a very high frequency of presentation. In Figure 4.2, the reference distribution, influenced by different α, parameters is shown in a logarithmic scale. Here, we can more appreciate the effects, in popularity, of α.

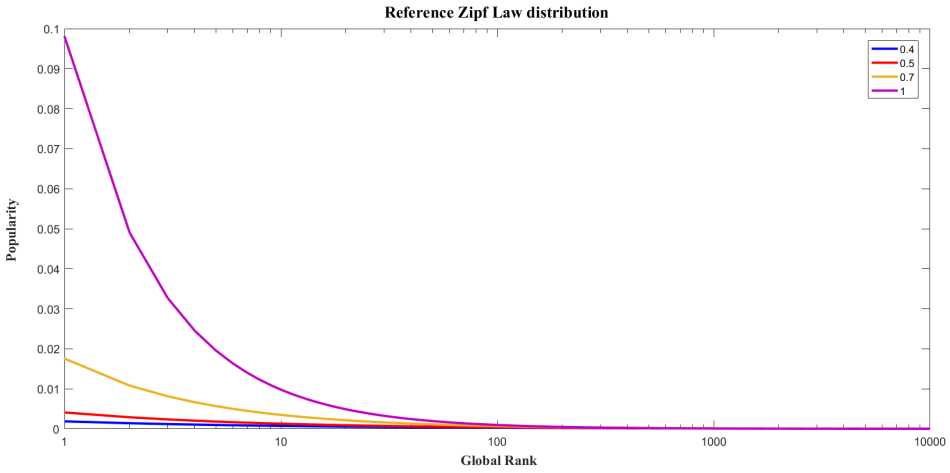

Figure 4.2: Reference Distribution affected by different α parameter in logarithmic scale

The reference distribution is used as the global popularity distribution, since the latest is not available in the first step.

4.2 Local Popularity Distributions

Once the reference distribution is available, it is possible to think concretely how to generate local popularities' distributions, having not mathematical models. In the thesis, this issue is treated in an empirical way starting from the reference popularity and supposing local popularity will follow, in average, that distribution. Therefore, in the local distribution the reference trend cannot be completely twisted, meaning, in average, files very popular have to be quite popular locally as well. This is related to the fact that local distributions are correlated to the global trend: even if locally, somewhere, some file can have a frequency very different from the one it has globally, in average it will be requested more or less with the same frequency. This means a file very popular globally, with high probability, will be very popular also locally even if with different values. The file rank changes but not significantly and it is quite unlikely the most popular file globally becomes the less popular in a lot of E-BSs. In fact, based on this, file ranks and popularity values can be different from the reference ones but the local distribution has to be carefully generated, in order to not revolutionize the real sense of what is a popular file, with its frequency and its priority in caching. In order to do this, a permutation lenght parameter, p, is defined. It is the length used to permute the vector made by reference ranks, which will be permuted step by step considering sub-vectors with permutation lenght to guarantee permuted ranks will be **exchanged with**

adjacent ranks, around their position index. In this way, locally, the importance of a file is changing maintaining its correlation with the reference rank, meaning p is directly related to the correlation between global distribution and the local popularity. The permutation is semi-random and it is done changing only files' ranks, not affecting values of popularity. Once the local rank vector is generated, each file has a popular value which was had by the previous file occupying that position, in the reference distribution. For instance if the file with rank 1 is exchanged with the file having rank 3, the file which locally has rank 3 has as popular value the popularity had by the file with rank 1 in the reference distribution therefore at the considered cache , the file with rank 3 is the most popular. This means, **locally, files don't have popularity according to the Zipf-law because that value is not directly derivable applying the law to the file rank**.

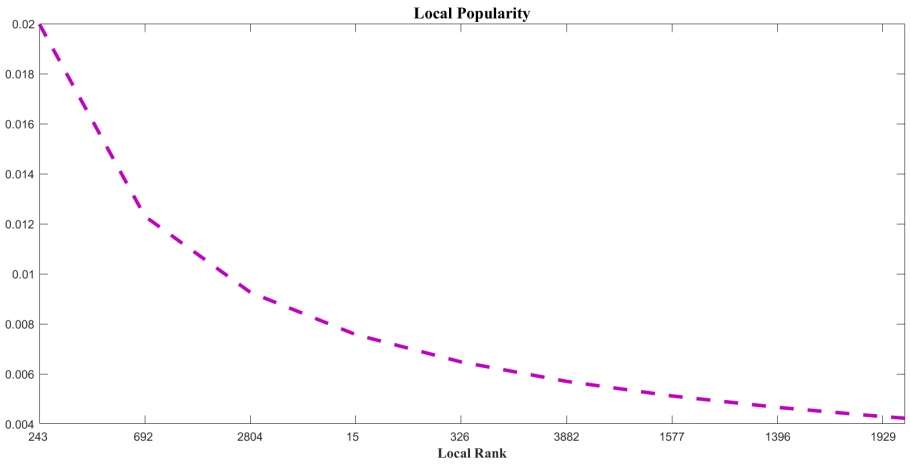

Figure 4.3: Local distribution with not sorted ranks in logarithmic scale

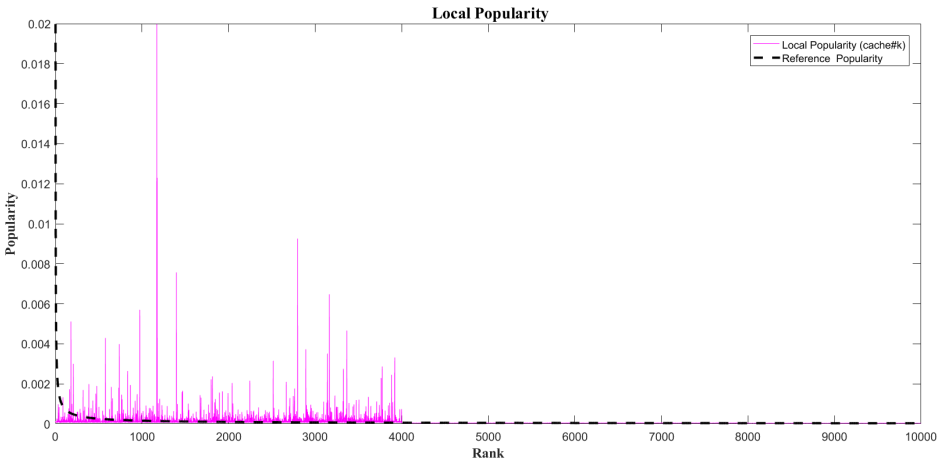

Figure 4.4: Local popularity distribution

In order to show results, Figure: 4.3 presents the popularity shape. It is the same of the reference one but rank , are different. The file with rank 11, for example, is, in the considered E-BS, the most popular locally. The local trend is represented in logarithmic scale. In Figure: 4.4, instead, ranks are sorted from 1 to N. The popularity shape is significantly different from the reference one. Obviously, all this is done at each E-BS in order to know local distributions everywhere in the network. The replicated scenario is quite realistic: some files are somewhere more popular than in other locations in the network but at the same time not totally decorrelated from the global popularity description of that file. To propose a consistent example, we suppose the most requested files are related to weather forecasts. The great part of E-BSs will experiment those kind of contents are, more or less, quite popular everywhere. However, somewhere in the network people are not so interested in that content and their popularity is very low even if quite unlikely the less one. In terms of Matlab code this is reflected by the p therefore, since caching is based on popularities, p affects the hit ratio: as bigger p as lower the hit ratio, meaning reference and local distributions are not so correlated. When p is short, local distribution is like reference distribution, instead, when it is long, at maximum is equal to N, distributions are completely different. In Figure 4.4 the distribution of local contents is specified. Storing global contents in this way can be not useful in terms of that metric, since using high values of p, global trends do not affect local behaviours. In order to better understand p meaning and effects on system, Figures show different local trends, generated from the same reference distribution but applying different permutations.

4.3 Global Popularity

Based on local popularities, the global distribution is calculated as the average all over E-BSs, managed by the satellite in the network. The Figure 4.5 shows the trend of the computed popularity. The shape has the same trend of the reference distribution but popularity values change. To better appreciate differences, the computed global distribution is presented in a logarithmic scale.

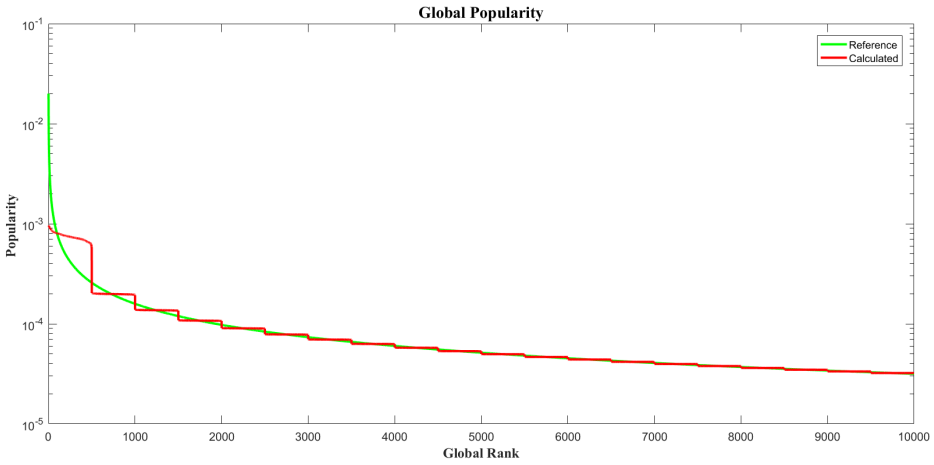

Figure 4.5: Computed global distribution

4.4 Placement Algorithm

The placement algorithm is the central point of our work in order to verify **the satellite actually has more advantages when it cooperates with the terrestrial network it lights up** and to measure benefits, in terms of system metrics, hit ratio and placement time. The local-only placement, storing contents according to local distributions is very efficient in terms of hit ratio. Instead, the global-only placement, storing contents only according to global distribution, is very efficient in terms of time placement. The hybrid placement is created as a mix of local-only and global-only placements. The research of an optimal hybrid placement aims to find a method of storing files allowing the maximum hit ratio and the minimum placement time. A trade-off between metrics is findable, based on the application and on the QoE requested at final user. Using Matlab, the function *PlacementAlgorithm* realizes the three different considered placements and gives back matrix containing saved files per each E-BS, using both methods.

The implemented algorithm works as follows:

Data:
Global Popularity
Local Popularity
C, Cache size
Threshold
Global Rank
Local Rank
while *there is space in the cache* **do**
    The cache is filled by files, according to the three selected algorithms;
end
Users' Requests Generation;
Hit Ratio Calculation;
Time Placement Calculation;
Result:
Stored files in cache using local-only placement
Stored files in cache using global-only placement
Stored files in cache using hybrid placement

Algorithm 2: Placement Algorithm

4.4.1 Local-Only Placement

The placement algorithm based only on local contents, stores first C files having the highest popularities. Reordering data according to the popularity values, the shape is the same of the reference one but looking at files' ranks it can be seen stored files are not, necessarily, having the rank in [1 ,C]. Sorting data based on the ranks, the popularity of stored contents is perfectly the same of the first C files in the local distribution at that E-BS.
Figure 4.6 shows popularity of stored files, in a generic edge cache, according to local-only placement. We can see it is exactly the same of first C files in local distribution.

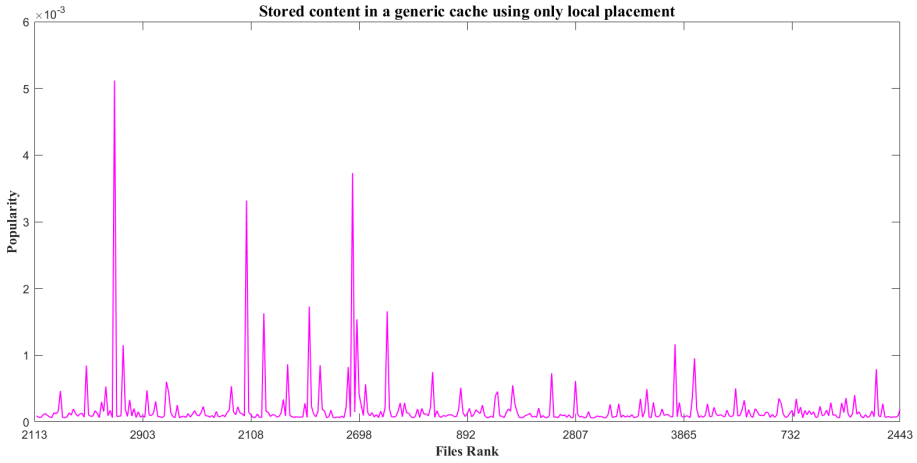

Figure 4.6: Popularity distribution of stored files using local placement

The created function gives as output stored files, identified by their ranks, to which a local popularity value is assigned.

4.4.2 Global-Only Placement

In the global placement algorithm only globally popular contents are stored in caches. Obviously, this means first C most popular contents, globally, are stored. Looking at popularities of in-cache files, they create the global popularity distribution of first C files, following the Zipf-like law. The popularity if stored file is shown in Figure 4.7, using a logarithmic scale.

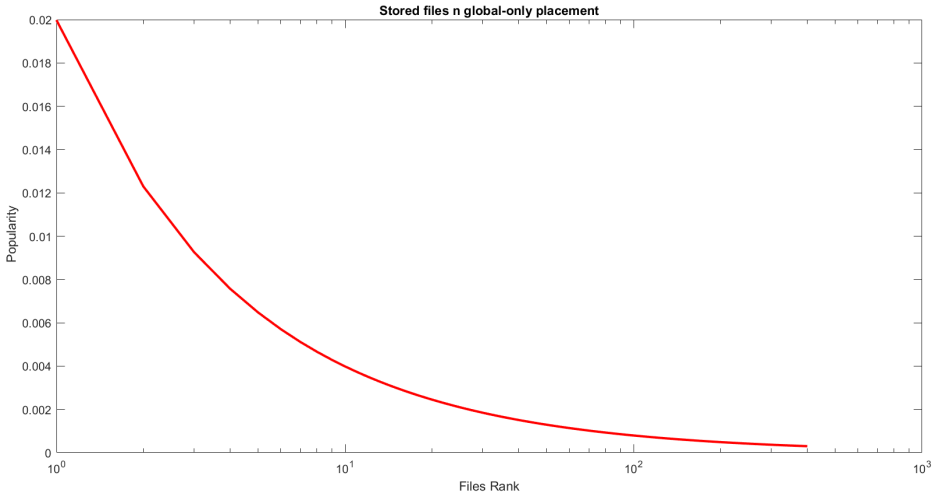

Figure 4.7: Popularity distribution of stored files in global-only placement(log scale)

4.4.3 Hybrid Placement

Two different hybrid approaches are treated: the first one is based on files rank and the second one based on files effective popularity values.

Centralized Hybrid Placement

This method is characterized by the fact that all caches have the same amount of global stored contents and the same amount of local stored contents, at esch E-BS.has been more developed in the thesis project and starting from that a lot of observations can be done based on so-obtained results. In the first step, a threshold th is fixed as:

$$th = \frac{CacheSize}{2} \tag{4.1}$$

First stored files are the first th files having rank in the range $[1, th]$ with higher probabilities. The second part of the cache, instead, is filled referring to the local distribution: first th ranks are taken from their not sorted distribution since they are the first more popular files, locally. If them or some of them have been already stored in the first part, the algorithm goes forward storing other files, since the cache is filled. The function gives back stored files' ranks and a matrix in which popularity values associated to those stored files, stored in the caches. Observing trends of in-cache contents, the first th reproduce exactly their global trend and in the second part more or less the local trend. The latest in fact can be a bit different because it is possible some files are not stored, since their are already taken in the first part of the algorithm and so they are not represented any more. Popularity of stored contents is shown, in logarithmic scale, in Figure 4.8 .

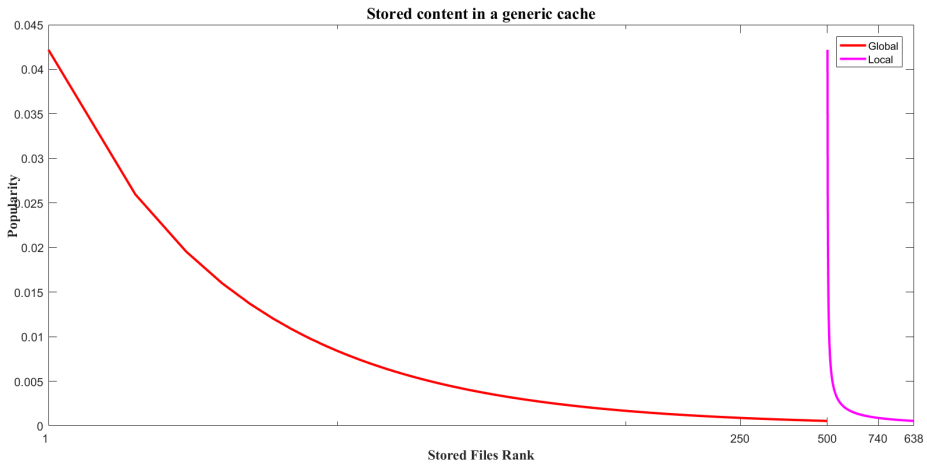

Figure 4.8: Stored contents in a generic cache (log scale)

Optimal threshold

The optimal threshold is defined as the threshold used in caching to discriminate the amount of global and local contents to store, in order to improve performance. The research of this threshold is investigated in an empiric way, varying the threshold from one to the *CacheSize*, since a mathematical model for it doesn't exists, not yet al least. In that manner, per each considered threshold, hit ratio and placement time have been measured allowing to find **the optimal working point** in which the two considered metrics belong to a given range values, which is defined fixing a trade-off between hit ratio and placement time.

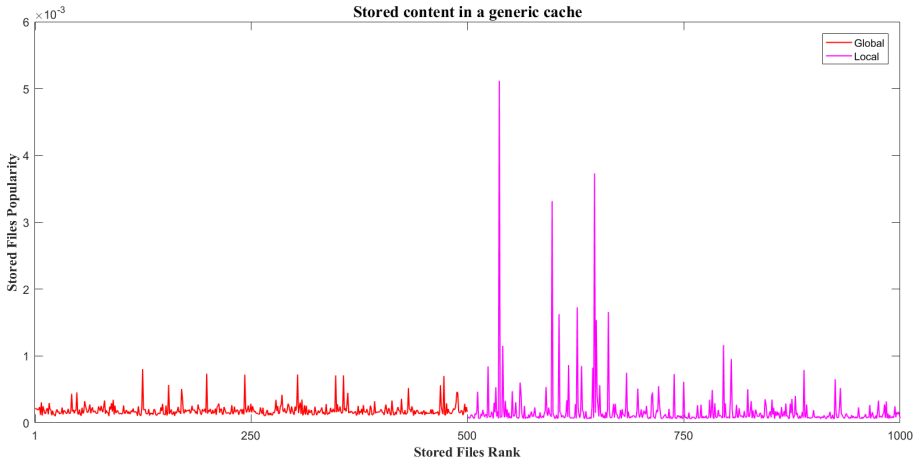

Figure 4.9: Stored contents in a generic cache with not sorted popularity values

4.4.4 Decentralized Hybrid Placement

The second investigated method doesn't rotate around ranks, related to the presentation frequency, although it focuses on the popularity value taken on each file. The main idea is to find an adaptive way to store contents in caches instead of a fixed threshold, without verifying rejected files are important in terms of popularity compared to the stored ones. An average value of global popularity is calculated averaging popularity values obtained in the computed global distribution. This value, Ga becomes the discriminant in storing contents. Starting from the local distribution, the local popularity value is considered per each file: if it is higher than Ga it is stored in the edge cache, if it is not it is rejected. This continues until when the cache is full. If local files having higher popularity values are not enough to fill the cache, the global distribution is considered. Global contents with higher popularity values can be stored in the cache, if not already present, until there is not space in the cache any more. Referring to two different caches, the stored files'

popularity trend is shown in Figure 4.10:

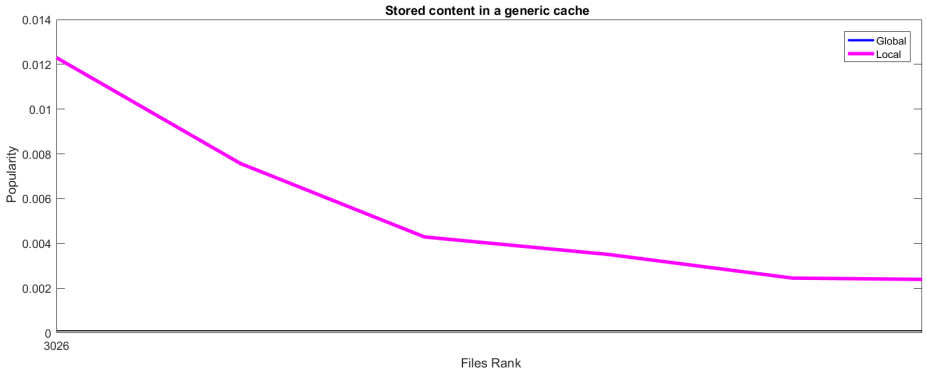

Figure 4.10: Stored contents in a generic cache using decentralized hybrid placement

With this approach, different caches store very different contents. It can be easily supposed this approach is better in terms of hit ratio since local distribution has more affect than the global one but, exactly because of this, performance in terms of time decreases since they have to use unicast links.

4.5 Users' requests generation

We suppose the number of requests, NoR, is the same in each E-BS. Local requests distribution is related to the popularity distribution in the cell, meaning users request contents according to the local popularity of files. Based on this, in order to generate users' requests we generate random requests based on the local distribution, obtaining the Figure 4.11. We can see users requests actually have a trend following local popularity distribution at that cache.

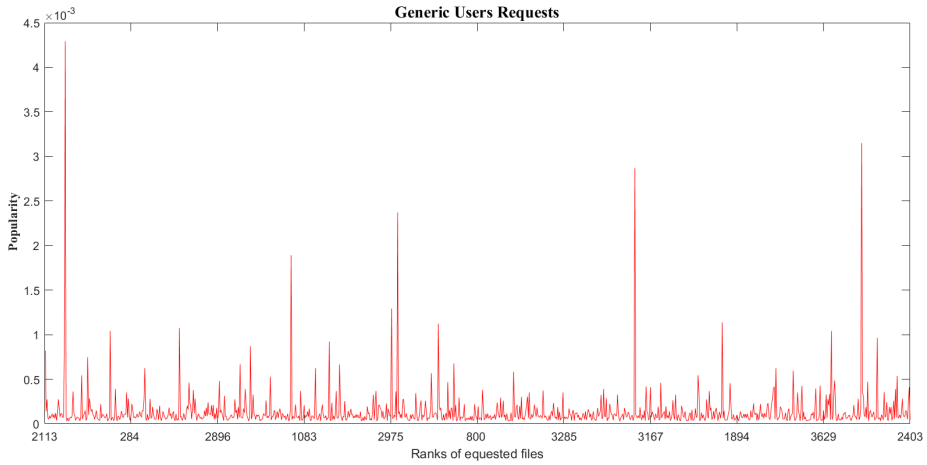

Figure 4.11: Users' requests in a generic cell(popularity shape)

4.6 Hit Ratio

The most important metric we consider to evaluate placement performance is the hit ratio, which is defined as:

$$HR = \frac{V}{NoR} \tag{4.2}$$

Where V is the number of in-cache requests. Since requests are generated starting from local distributions, when only local popular contents are stored HR is maximum. It is minimum when only globally popular contents are saved. In Figure 4.12 it is shown the comparison among the hit ratios obtained using the global, local and hybrid placements. The hybrid placement has very good performance, also when the cache size is very limited cache size, if compared to library size. In fact, hybrid hit ratio is not so far from performance achieved using the local-only placement, which is the upper-bound. Wors performance are related to the global-only approach.

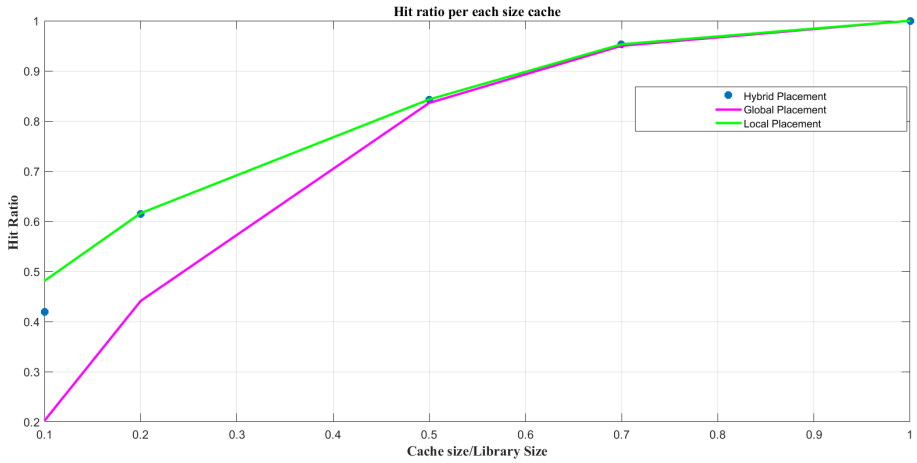

Figure 4.12: Hit ratio comparison using the centralized approaches:global local and hybrid

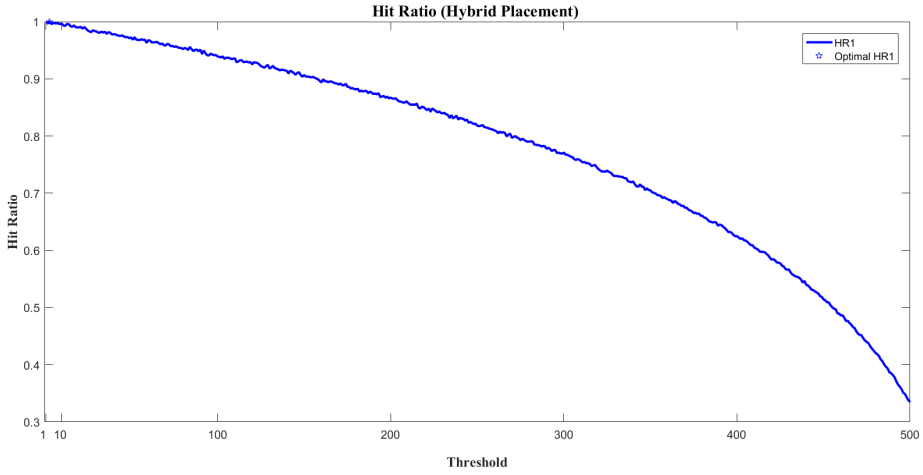

Figure 4.13: Hit ratio using hybrid placement

In Figure 4.13, hit ratio trend is shown, varying the threshold. In Figure 4.14, hit ratios obtained using centralized and decentralized approach can be seen. They are presented varying the threshold, discriminating local and global contents to store in the centralized placements. Actually, this threshold doesn't affect the second placement, which focus on popularities values considerations. They are presented in the same figure to show decentralized placement has performance more or less constants. This happens since the algorithm considers, cache by cache, contents to store, avoiding to lose local informations if they are more important compared to the global one.

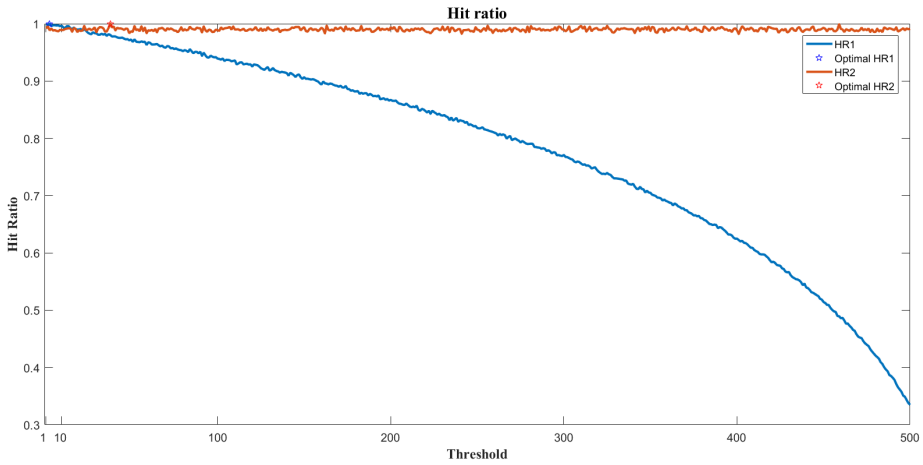

Figure 4.14: Hit ratio comparison using centralized and decentralized approaches

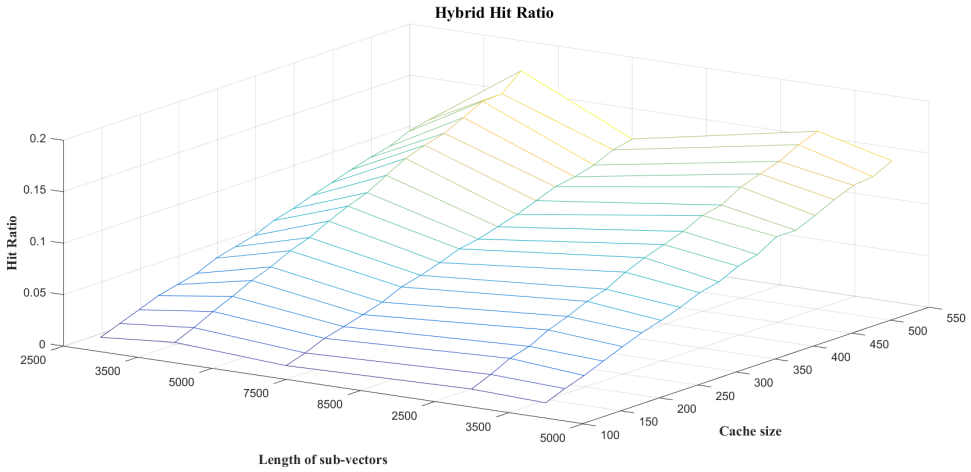

Figure 4.15: Hit Ratio 3-D

In Figure 4.15 a 3-D representation of hit ratio is shown, in order to take in count this metric depends, not only on cache size or placement algorithms but it is strongly related on the way of generation local popularities. Therefore hit ratio is affected also by p, the length of sub-vectors using in permutation to generate local distributions. Also the multibeam scenario is investigated, achieving important performance in terms of hit ratio. It is shown in Figure 4.16, considering also different α parameters. The multibeam improves performance since the global popularity is referred to the beam, considering the sub-network.

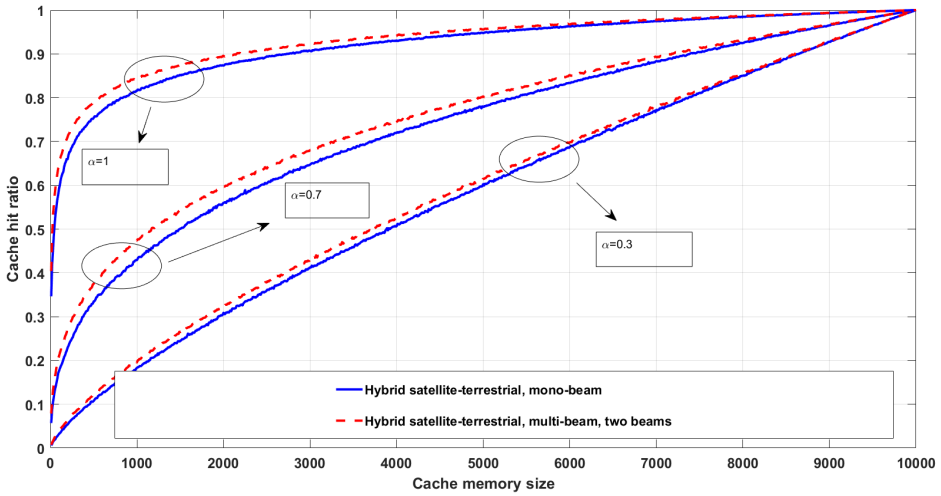

Figure 4.16: Hit ratio comparison using monobeam and multibeam satellite

4.7 Time Placement

The second evaluated metric is the placement time, which refers to the time needed to fill the caches. The terrestrial links uses unicast transmission instead of the satellite, which uses multicast or broadcast. They are defined as follows:

- **Satellite placement time**

$$TPG = \frac{F}{Backhaul\_rate} \tag{4.3}$$

- **Terrestrial placement time**

$$TPL = \frac{F * K}{Backhaul\_rate} \tag{4.4}$$

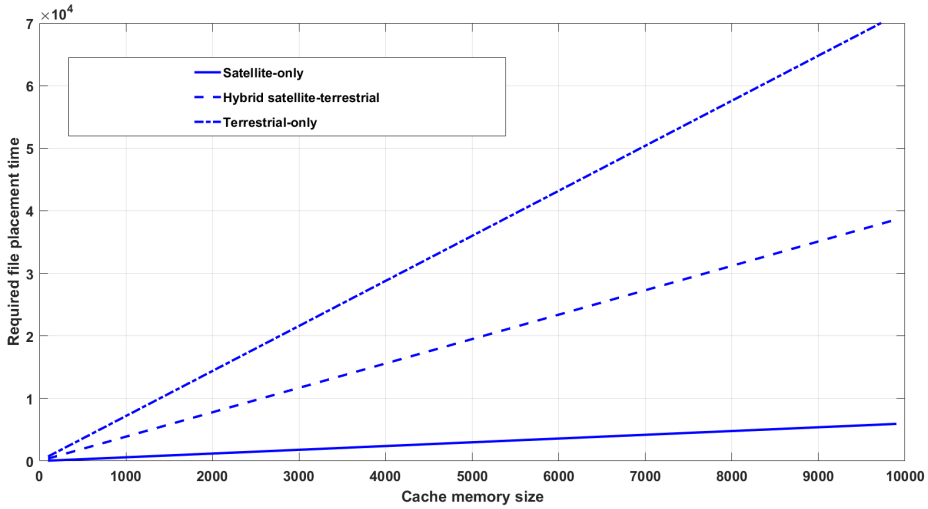

Figure 4.17: Placement time

Where F is the number of files to send and K is the number of caches. In Figure 4.17 it is shown what happens using local-only and global-only and hybrid placements. The time needed to fill the caches increases in any case but the slope is very different: when caches have a lot of memory, satellite-only placement is not affected by the fact a hug amount of data have to be sent. This doesn't happen in the local-only approach, in which time increases a lot while the cache size increases. In the medium there is the proposed hybrid placement, which has a slope higher than the only-global approach and slower than the only-local approach.

$$TP_TOT = TPG + TPL \tag{4.5}$$

Chapter 5

Future Works

Satellite-assisted caching in hybrid networks is very advantageous, in terms of system performance and users QoE. Applying coded caching, some benefits can be obtained in satellite communications:

- different files can be transmitted to the users via the broadcast stream of a mono-beam satellite.
- users in different geographical locations, e.g., different countries, may demand different types of files. Users in a specific geographical location demand similar files so that they can be served by a multibeam satellite, where each beam targets a specific geographic location. Here, coed caching strategy can be used for providing, via multicast, more than on file to the users of a given geographical location, increasing the performance of a multibeam satellite

Moreover, a lot of scenarios could benefit from satellite-assisted caching such as high altitude platforms (HPAs)-enabled networks (e.g. Google Loon Project, Facebook's Connectivity Lab). HAPs are equipped with caches that can be fed jointly by terrestrial and satellite links. Aeronautical caching for in-flight entertainment, with satellite-assisted caching can make a difference. In this case, the airplanes are equipped with caches to facilitate the user experience while streaming contents. The airplanes could be connected directly to the ground or by regional satellite beams.

Chapter 6

Conclusions

This thesis focused on hybrid networks, composed by a satellite overlay over terrestrial content delivery networks. Current CDNs aim to store contents at the edge of the network, at E-BSs, according to files' popularity. E-BSs can compute local popularities since they have a local view of the network, taking into account local users requests. The satellite, thanks to its high coverage area, can learn local popularities and, based on them, can calculate the global popularity distribution. Since local users requests are affected by global trends, in hybrid networks, satellite-assisted caching is realized storing contents based on, jointly, local and global popularities of contents.

In this work, we compared different placement algorithms: *local-only based*, *global-only based* and *the implemented hybrid placement*. The first two paradigms store most popular contents according to local and global popularity distributions, respectively. In the local-only based strategy, contents are sent to the caches via unicast, on the contrary, in the global-only based approach the transmission is multicast/broadcast, according to the employed satellite (multibeam or monobeam). Performance of both placement algorithms are used as baseline: the first one is the best in terms of hit ratio, the second one is the best in terms of placement time. We focused on hybrid placement approach implementation, which aims to store contents in the best possible way, using both local and global popularities, combining unicast and multicast/broadcast transmissions. According to a fixed threshold, a certain amount of globally popular contents is stored. The remaining memory is filled using locally popular data. We investigated two different approaches to fix the threshold: the centralized approach, in which the threshold is the same for each edge cache, and the decentralized approach in which the threshold is determined cache by cache. The hybrid strategy is implemented both in monobeam and in multibeam satellite overlay, achieving interesting results. Despite the local-

only placement, it shows better performance in terms of time placement. In the multibeam satellite scenario, the hybrid placement achieves higher performance, compared to the monobeam, since the global popularity distribution is more accurately computed. Presented results allow to define a trade-off between hit ratio and placement time, based on users requirements. Hit ratio, using the hybrid placement, approximately, has the same performance of the local-only based approach, even when the cache size is significantly limited. In the placement time, instead, hybrid performs much better than the local-only approach. Supposing the number of files to send and the backhaul rate are both the same for terrestrial and satellite networks, and fixing the cache size, the hybrid based placement time needs two times less than the local-only based placement. Focusing on popular contents the algorithm, indeed, realizes smart caching at the edge, minimizing the transmissions by terrestrial backhaul and creating a users content-oriented network. The hybrid network architecture along with the proposed hybrid caching algorithm, improve system performance in terms of users QoE and bandwidth consumption.

Appendix A

Hybrid system metrics

In order to evaluate the system performance, several metrics have been proposed. Hereafter, some of the most important ones are presented [12], [34]:

- Cache hit ratio.
  It is the ratio of the number of requests stored in the cache and the total number of requests and it is used to evaluate the selection efficiency of cached content. In the case of multiple caches, this metric is derived by averaging over multiple cache hit ratios.

- Latency time.
  This is the delay between a file request and file receiving. Obviously, when it is too long users requirements cannot be satisfied and performance decreases. If the file is cached, the effective delay is shorter because the content is available in just one hop.

- Hybrid backhaul traffic volume
  Two different definitions of this metric are presented, based on the implemented type of caching algorithm:

  1. online: The metric refers to the traffic imposed over the terrestrial backhaul due to the absence of the requested files in the cache, which must be routed through the core network.
  2. offline: it is the amount of traffic passing through the terrestrial backhaul link in both placement and delivery phase, characterizing offline caching.

- terrestrial backhaul traffic volume
  This metric describes the amount of data exchanged by the CDN's nodes and

the core network. It is the ratio of total byte transferred through this link and the total number of requests.

- Cache-feeding Throughput
  This metric is used to evaluate the efficiency of cache feeding. It can be measured by measuring the time needed for placing a predetermined data volume across the system caches. It is the time needed for placing a predetermined data volume across the system caches.

- User Throughput
  It is the effective data rate that the user perceives when requesting a file.

Bibliography

[1] Chris Brinton, Ehsan Aryafar, Steve Corda, Stan Russo, Ramiro Reinoso, and Mung Chiang. An intelligent satellite multicast and caching overlay for cdns to improve performance in video applications. In *31st AIAA International Communications Satellite Systems Conference*, page 5664, 2013.

[2] Ivanes Lian Costa Araujo and Aldebaro Klautau. Traffic-aware sleep mode algorithm for 5g networks. In *Telecommunications (IWT), 2015 International Workshop on*, pages 1–5. IEEE, 2015.

[3] Pavan Kamaraju. Towards content delivery optimization in future wireless networks. In *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2016 IEEE 17th International Symposium on A*, pages 1–3. IEEE, 2016.

[4] Xi Peng, Juei-Chin Shen, Jun Zhang, and Khaled B Letaief. Backhaul-aware caching placement for wireless networks. In *2015 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE, 2015.

[5] Amine Abidi and SoniaMettali Gammar. Towards new caching strategy for information-centric networking based on data proximity control. In *Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (CIT/IUCC/DASC/PICOM), 2015 IEEE International Conference on*, pages 540–547. IEEE, 2015.

[6] Peter Hillmann, Tobias Uhlig, Gabi Dreo Rodosek, and Oliver Rose. Modeling the location selection of mirror servers in content delivery networks. In *Big Data (BigData Congress), 2016 IEEE International Congress on*, pages 438–445. IEEE, 2016.

[7] Vitor Jesus and Rui L Aguiar. Figures of merit for the placement (in) efficiency of interconnected cdns. In *Computers and Communications (ISCC), 2012 IEEE Symposium on*, pages 000277–000282. IEEE, 2012.

[8] Rajeev Tiwari and Neeraj Kumar. A novel hybrid approach for web caching. In *Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2012 Sixth International Conference on*, pages 512–517. IEEE, 2012.

[9] Hatem Ibn-Khedher, Emad Abd-Elrahman, and Hossam Afifi. Omac: Optimal migration algorithm for virtual cdn. In *Telecommunications (ICT), 2016 23rd International Conference on*, pages 1–6. IEEE, 2016.

[10] M Zubair Shafiq, Amir R Khakpour, and Alex X Liu. Characterizing caching workload of a large commercial content delivery network.

[11] Luigi Rizzo and Lorenzo Vicisano. Replacement policies for a proxy cache. *IEEE/ACM Transactions on Networking (ToN)*, 8(2):158–170, 2000.

[12] TR Nair and P Jayarekha. A rank based replacement policy for multimedia server cache using zipf-like law. *arXiv preprint arXiv:1003.4062*, 2010.

[13] P Venkataram, Shashikant Chaudhari, R Rajavelsamy, TR Ramamohan, and H Ramakrishna. Disk-oriented vcr operations for a multiuser vod system. *Journal of the Indian Institute of Science*, 84(5):123, 2013.

[14] Zhan-sheng Li, Da-wei Liu, and Hui-juan Bi. Crfp: A novel adaptive replacement policy combined the lru and lfu policies. In *Computer and Information Technology Workshops, 2008. CIT Workshops 2008. IEEE 8th International Conference on*, pages 72–79. IEEE, 2008.

[15] Nimrod Megiddo and Dharmendra S Modha. Arc: A self-tuning, low overhead replacement cache. In *FAST*, volume 3, pages 115–130, 2003.

[16] Mudashiru Busari and Carey Williamson. On the sensitivity of web proxy cache performance to workload characteristics. In *INFOCOM 2001. Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, volume 3, pages 1225–1234. IEEE, 2001.

[17] Lee Breslau, Pei Cao, Li Fan, Graham Phillips, and Scott Shenker. Web caching and zipf-like distributions: Evidence and implications. In *INFOCOM'99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, volume 1, pages 126–134. IEEE, 1999.

[18] Elizabeth J O'neil, Patrick E O'neil, and Gerhard Weikum. The lru-k page replacement algorithm for database disk buffering. *ACM SIGMOD Record*, 22(2):297–306, 1993.

[19] Hasti Ahlehagh and Sujit Dey. Video-aware scheduling and caching in the radio access network. *IEEE/ACM Transactions on Networking (TON)*, 22(5):1444–1462, 2014.

[20] Ejder Baştuğ, Mehdi Bennis, Engin Zeydan, Manhal Abdel Kader, Ilyas Alper Karatepe, Ahmet Salih Er, and Mérouane Debbah. Big data meets telcos: A proactive caching perspective. *Journal of Communications and Networks*, 17(6):549–557, 2015.

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

[22] Mohammed S ElBamby, Mehdi Bennis, Walid Saad, and Matti Latva-Aho. Content-aware user clustering and caching in wireless small cell networks. In *2014 11th International Symposium on Wireless Communications Systems (ISWCS)*, pages 945–949. IEEE, 2014.

[23] Maryan Kyryk, Nazar Pleskanka, and Maryana Pitsyk. Qos mechanism in content delivery network. In *2016 13th International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science (TCSET)*, pages 658–660. IEEE, 2016.

[24] Hilmar Linder, Horst D Clausen, and Bernhard Collini-Nocker. Satellite internet services using dvb/mpeg-2 and multicast web caching. *IEEE Communications Magazine*, 38(6):156–161, 2000.

[25] Pablo Rodriguez and Ernst W Biersack. Bringing the web to the network edge: large caches and satellite distribution. *Mobile Networks and Applications*, 7(1):67–78, 2002.

[26] Nicolas Courville, Hermann Bischl, Erich Lutz, Ales Svigelj, Pauline ML Chan, Evangelos Papapetrou, and Rafael Asorey-Cacheda. Hybrid satellite/terrestrial networks: State of the art and future perspectives. In *QShine 2007 Workshop: Satellite/Terrestrial Interworking*, page 1. ACM, 2007.

[27] ETSI Satellite Earth Stations. Systems (ses); broadband satellite multimedia (bsm). *Interworking with IntServ QoS*, 2007.

[28] Miguel Ángel Vázquez, Luis Blanco, Xavier Artiga, and Ana Pérez-Neira. Hybrid analog-digital transmit beamforming for spectrum sharing satellite-

terrestrial systems. In *Signal Processing Advances in Wireless Communications (SPAWC), 2016 IEEE 17th International Workshop on*, pages 1–5. IEEE, 2016.

[29] Symeon Chatzinotas, Bjorn Ottersten, and Riccardo De Gaudenzi. *Cooperative and Cognitive Satellite Systems*. Academic Press, 2015.

[30] Fabián Mendoza, Ramon Ferrús, and Oriol Sallent. Flexible capacity and traffic management for hybrid satellite-terrestrial mobile backhauling networks. In *Wireless Communication Systems (ISWCS), 2016 International Symposium on*, pages 119–124. IEEE, 2016.

[31] Xiaoming Zhou and John S Baras. Tcp over geo satellite hybrid networks. In *MILCOM 2002. Proceedings*, volume 1, pages 29–34. IEEE, 2002.

[32] Jing Zhu, Sumit Roy, and Jae H Kim. Performance modelling of tcp enhancements in terrestrial-satellite hybrid networks. *IEEE/ACM Transactions on Networking (TON)*, 14(4):753–766, 2006.

[33] Jiaxin Zhang, Barry Evans, Muhammad Ali Imran, Xing Zhang, and Wenbo Wang. Green hybrid satellite terrestrial networks: Fundamental trade-off analysis. In *Vehicular Technology Conference (VTC Spring), 2016 IEEE 83rd*, pages 1–5. IEEE, 2016.

[34] Aner Armon and Hanoch Levy. Cache satellite distribution systems: Modeling and analysis. In *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*, volume 1, pages 240–250. IEEE, 2003.

[35] Sameh Sorour and Shahrokh Valaee. A network coded arq protocol for broadcast streaming over hybrid satellite systems. In *2009 IEEE 20th International Symposium on Personal, Indoor and Mobile Radio Communications*, pages 1098–1102. IEEE, 2009.

[36] Jiaxin Zhang, Barry Evans, Muhammad Ali Imran, Xing Zhang, and Wenbo Wang. Performance analysis of c/u split hybrid satellite terrestrial network for 5g systems. In *Computer Aided Modelling and Design of Communication Links and Networks (CAMAD), 2015 IEEE 20th International Workshop on*, pages 97–102. IEEE, 2015.

[37] Edward J Nossen. Satellite communications system, February 21 1995. US Patent 5,392,450.

[38] Nedo Celandroni, Erina Ferro, Giovanni Giambene, and Mario Marandola. Sat01-3: Tcp performance in a hybrid satellite network by using acm and arq. In *IEEE Globecom 2006*, pages 1–6. IEEE, 2006.

[39] Mei Li and John J Metzner. Reliable satellite multicast with assistance of terrestrial communications. In *Circuits and Systems, 2002. MWSCAS-2002. The 2002 45th Midwest Symposium on*, volume 1, pages I–396. IEEE, 2002.

[40] Giuliana Iapichino, Christian Bonnet, Oscar del Rio Herrero, Cedric Baudoin, and Isabelle Buret. Advanced hybrid satellite and terrestrial system architecture for emergency mobile communications. In *Proc. 26th AIAA International Communications Satellite Systems Conference (ICSSC 2008)*, 2008.

[41] Mohammad Mohammadi Amiri, Qianqian Yang, and Deniz Gunduz. Coded caching for a large number of users. *arXiv preprint arXiv:1605.01993*, 2016.

[42] Sheng Yang, Khac-Hoang Ngo, and Mari Kobayashi. Content delivery with coded caching and massive mimo in 5g. In *Turbo Codes and Iterative Information Processing (ISTC), 2016 9th International Symposium on*, pages 370–374. IEEE, 2016.