Enhanced Caching Strategies At the Edge of LTE Mobile Networks

Andre S. Gomes *†, Torsten Braun *, Edmundo Monteiro †
*Institute of Computer Science, University of Bern, Switzerland
{gomes,braun}@inf.unibe.ch
†CISUC, University of Coimbra, Portugal
{asng,edmundo}@dei.uc.pt

***Abstract*—With a boom in the usage of mobile devices for traffic-heavy applications, mobile networks struggle to deliver good performance while saving resources to support more users and save on costs. In this paper, we propose enhanced strategies for the preemptive migration of content stored in Information-Centric Networking caches at the edge of LTE mobile networks. With such strategies, the concept of content following the users interested in it becomes a reality and content within caches is more optimized towards the requests of nearby users. Results show that the strategies are feasible, efficient and, when compared to default caching strategies, ensure that content is delivered faster to end users while using bandwidth and storage resources more efficiently at the core of the network.**

***Index Terms*—Information-Centric Networking, Content Migration, Caching, LTE.**

I. Introduction

Mobile network evolution in the last few years has been quite intense, with major increase of throughput performance and resources usage efficiency. Such evolution is mostly driven by tremendous demand of bandwidth [1], on the one hand because smartphones and other mobile devices play a major role as content demanders, and on the other hand because traffic-heavy applications are part of the daily life of millions of people. However, satisfying the content requirements of the current number of users with such dynamic networks is still an open challenge, which is currently being addressed by a number of emerging concepts and technologies.

As far as the network is concerned, new 5G concepts such as Network Function Virtualization (NFV) [2] are emerging, allowing mobile networks to adapt more dynamically to different conditions and requirements, and also to support other value-added technologies. One of these efforts is Cloud Radio Access Network (C-RAN) [3][4]. It brings the possibility to virtualize the entire 3GPP Long Term Evolution (LTE) radio infrastructure, except for the antennas. Virtualized infrastructures extend the cloud computing concept to the Radio Access Network (RAN), and explore the modularity of the components together with the usage of general-purpose hardware infrastructure to run evolved Node Bs (eNBs). Such fact transforms C-RAN into an enabler for deployment of value-added services closer to the edge of mobile networks, i.e. in very close proximity to mobile users. Despite increased delays due to its characteristics, the proximity deployment of other services allows for performance gains and cost savings that more than surpass those overheads and improve the end-to-end service.

ISBN 978-3-901882-83-8 © 2016 IFIP

In this direction, Future Internet (FI) concepts such as Information-Centric Networking (ICN) [5], which proposes a change in the current host-centric paradigm of requesting content, are becoming increasingly important due to the advantages brought together by its content-centric architecture. Namely: performance improvements [6], indirect bandwidth savings from its caching-based architecture, enhanced mobility support [7] and increased security [8].

With such concepts in mind, proposals appear to take advantage of the fact that they complement each other. Gomes et al. [9] evaluate the feasibility of deploying ICN together with 3GPP LTE mobile networks, leveraging the C-RAN concept and its role as an enabler for the deployment of additional services at the edge of these mobile networks. In that work, authors conclude that there are clear benefits of deploying ICN routers co-located with LTE eNBs, such as bandwidth savings at the core network and lower latency to retrieve content derived from the proximity to end users. Those findings are also in line with works such as [6], and show that there is an important demand of enhanced caching strategies to have content cached closer to the users interest in it while using resources efficiently.

Those caching strategies are twofold: first they are used to populate edge caches, and thereafter they must maintain content where it will yield the most benefit at any given time. As users are increasingly mobile and tend to move between different locations quite often, it is safe to assume that what is cached at a location is not necessarily what is going to be requested by the users that will be there in the next few hours or days. Studies [10] even show that user interests in social media content contribute deeply to its locality and homophily characteristics, which means that people geographically close to each other may have common or similar interests of content objects (locality) [11] and also that users are clustered by regions and interests (homophily) [12]. That leads to the question: how should caching strategies handle user mobility? Such question does not have a simple answer, as some assumptions have to be made and challenges need to be considered to reach a preliminary conclusion. First, it is important that user mobility is predicted to perform preemptive actions, and proposals

exist to deal with it [13][14][15]. Then, if a set of users is predicted to be at a location, more complex decisions need to be made in order to have accurate migrations that minimize overhead and maximize performance. The first decision is whether content from caches at the origin of the users should be migrated to other caches at the possible destinations. Once that is established, other questions arise: where should content be migrated to (mobility prediction usually outputs a list of possible destinations with different probabilities), which subset of the content should be migrated, how it should be migrated and when should it be migrated.

In this paper, we attempt to answer the previously described questions by developing content migration strategies that handle the required decisions and deliver the greatest possible trade-off between benefit and cost. In section II, existing proposals to address content migration strategies are analyzed. Section III introduces our proposal for the migration of content and related decisions. Section IV describes experimentation scenarios for the evaluation of the proposal. Section V presents the results of the performed experiments. Finally, in section VI, the main achievements of this work are highlighted.

II. Related Work

When considering strategies that take into account the mobility of users to decide on placement/migration of services or content within mobile networks, only a few works exist and there are still a number of shortcomings to be addressed.

One proposal assumes that mobility of the user is considered for services placement and scaling [16]. In this case, orchestration of distributed cloud services is done by predicting user mobility, i.e. more or less resources are allocated if the system predicts that users will move to/from the location of each small Data Center (DC). However, migration of services from one location to another is not considered.

In this direction, one very important concept towards migration strategies - Follow-Me Cloud - was first proposed by Taleb et al. [17]. It essentially considers that small DCs are present closer to the edge of mobile networks and proposes that services are deployed in close geographical proximity to users. Hence, when users move to a different location, those services should be migrated and follow the user. To handle the decision, several different models can be used. An analytical model based on Markov Decision Processes is proposed [18][19].

Such model considers that user positions must be found in order to have services instantiated in the optimal DC. It relies on the random walk mobility model to try to predict future positions, and when the user is n hops away from its current optimal DC, migrations are triggered using a system modeled with Continuous-Time Markov Chains. Also, when considering if data migration should be done, factors such as class of the user, load policies, service migration costs and service migration duration need to be analyzed. Bearing these factors in mind, it is assumed that cost and service disruption are to be minimized, and the user should be connected to the optimal DC as often as possible. This approach provides many benefits for the migration of services, but only a final destination is considered, not multiple destinations along a path. Moreover, it does not decide which services to migrate, it only considers a single user (overhead of migration for a single user may not be justified) and does not deal with specificities of migrating content or even stateful services.

As far as strategies to deal with content migration are concerned, other works [20] have looked at the problem in Peer-to-Peer (P2P) networks from the provider perspective. With a typical hierarchical Content Delivery Network (CDN) architecture, the main requirement is to distribute content among nodes in a way that leaf nodes get the most traffic and root nodes are seldom used. This strategy increases performance and thus reduces latency for end users, and relies on decisions to migrate/copy content from one node to another depending on popularity and cost. However, as it maximizes the usage of caches while attempting to maximize performance, those decisions are a NP-complete problem that is hard to manage. In very dynamic mobile networks, that poses an issue due to the need of quick and proactive decisions sometimes even before there is user movement, and that is why other works aim at less complex and local approaches that maintain hierarchical caches efficiently used [21].

Another proposal [22] takes dynamic mobile networks into account, and uses proactive migration strategies for content, i.e. migration is triggered when it is predicted that the user will move to a neighbor location. Using a proxy system, it is proposed that subscribed content is pre-fetched whenever it is predicted that a user will move to the geographical region of another proxy. With the knowledge of possible destinations and corresponding probabilities, a decision has to be made in order to select the destination proxies for the content while minimizing cost (migration cost and cache storage) and maximizing benefit (latency and cache hit ratios). Despite some gains in terms of delay, the number of criteria for migration decisions is small and no different weights are considered, there are no replacement policies when caches become full and the required single user mobility prediction is too simplistic/naive, i.e. the effect of a single user on the entire network is questionable when comparing to the required overhead of content migration and cache usage.

Considering all the proposals and the issues they fail to address, we propose a system that relies on their positive findings and at the same time attempts to address the challenges not taken into consideration.

III. Enhanced Caching Strategies

In-line with the idea of Follow-Me Cloud (FMC) described in the previous section, the proposal can be summarized into making decisions and perform preemptive migrations of mobile network's edge-cached content based on user mobility, i.e. migrations (copies) ahead of future user requests at a new location. The following key objectives are assumed: mobility prediction must be used to take actions before users move from one location to another, content may only be migrated if it is likely it will yield benefit at the destinations, multiple

destinations may be considered at the same time to improve accuracy and migration cost should be minimized as much as possible. We also assume that content migrations should happen among caches with modified policies, as they typically implement a Least Recently Used (LRU) policy that can delete recently added content in a matter of seconds if the load of received Interests is high. In this case we are interested in maintaining our own policy, i.e. popularity based queue, with the most popular and recently accessed content at the head of the queue and deletions happening at the tail.

A. Architecture

In order to achieve the goals associated with the objectives of this proposal, an architecture was defined as illustrated in Fig. 1.

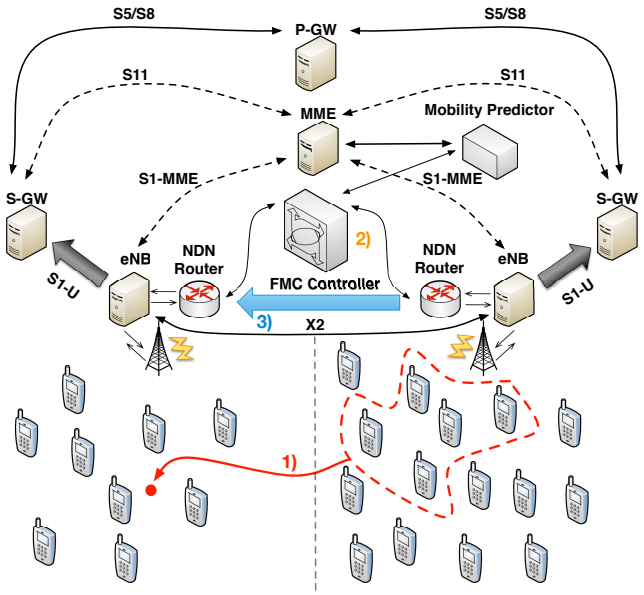

Fig. 1: Follow-Me Cloud Architecture

In this architecture, we assume as base architecture the 3GPP LTE Evolved Packet System (EPS). Namely, its main components (eNB, S-GW, P-GW, MME) and its main interfaces (X2, S1-U, S1-MME, S11, S5/S8). At the same time, a recent and increasingly popular approach for ICN is selected - Named Data Networking (NDN) [23]. We consider that NDN routers are co-located with 3GPP LTE eNBs [9], serving the subset of users present at each network cell. At the same time, information about the network and its users is gathered at the FMC Controller. This information includes mobility prediction data, local content popularity and availability at a given cell, number of users per cell and availability of resources such as storage. With that information, a multi-step process for content migration is triggered every time mobility is predicted or detected by the Mobility Predictor using input from the 3GPP LTE Evolved Packet Core's (EPC) Mobility Management Entity (MME), based on the defined time period between predictions:

1) FMC Controller is notified about user mobility (e.g. user ID n is moving from cell ID x to cell ID y with probability z) and decides, upon policies such as the number of users moving to a destination cell, if other steps should be taken or the process should stop.
2) If decisions have to be made regarding content migration, the FMC Controller has to decide: where to migrate content (cell ID and corresponding NDN router), what subset of the local content should be migrated (content object's prefixes), when to do it (according to other scheduled migrations) and finally how to do it (routing and load balancing for content requests).
3) After a decision is made, the FMC Controller issues a NDN message called Request of Interests that instructs the destinations' NDN router(s) to fetch the subset of content to be migrated from the closest source and place it at its cache.

B. Decision Techniques

As far as decisions are concerned, two types of decisions must be carefully analyzed and made in the proposed system. The first decision is **where** to migrate content when users are moving. Although mobility prediction will output a list of candidate destinations and respective probabilities, such set of destinations may not be complete and still needs to be reduced and ranked according to other important factors.

To increase the list of candidates, one may assume that neighbor cells in-between the returned destination candidates and the origin should also be considered. After all, the user will need to travel through those cells and, depending on the delta time, i.e. amount of time in the future considered for prediction, it may even stop and stay there longer than at the final destination. With this full list of destination candidates, the problem is now how to rank these in order to select only a few that satisfy the defined criteria and will yield the highest trade-off between benefit and cost.

Multiple-Criteria Decision-Making (MCDM) [24] is an approach to make decisions in the presence of multiple, usually conflicting, criteria. As it is not tied to a specific problem, it can be applied to a very diverse range of scenarios and problems, from business decisions to complex science problems. At the same time, it supports multiple weighted criteria and typically returns a finite number of solutions when dealing with a selection/assessment problem. Therefore, it fits the decision to be performed in terms of ranking/selection the destinations for content migration.

To handle decisions, which involve ranking of candidates, the most common methods are score methods. Within these, perhaps the most well-known and used method is Analytical Hierarchy Process (AHP) [25][26][27]. This method starts by summarizing the problem, deciding the hierarchical list of criteria to be considered for the decision and listing the alternatives to be ranked. In this case, the problem is already defined: a destination or destinations need to be selected for content migration. Considering that not only the destination of users is important but also to maximize the efficiency of cache usage within the network edge, the following criteria were defined:

- Mobility prediction information containing probabilities for destinations.
- Percentage of non-intersecting content between origin and alternative destination.
- Percentage of free storage space at alternative destination.
- Relative size of mobile group, defined as the ratio between number of users moving and users present at alternative destination.
- Cost of migration, estimated as the network transfer delay to copy the expected data size from its origin to an alternative destination.

Afterwards, a $N \times M$ matrix is created, where N is the number of alternatives and M the number of criteria. For each cell of the matrix, a score value is calculated to reflect how good the alternative is in terms of the criteria being considered.

As each of the criteria may have a different significance for the decision, each of them should also have a weight value to be considered. In order to rank criteria, judgment is used by creating a $M \times M$ matrix where each criteria is compared against the others using pair-wise comparisons, i.e. each compared to all the others in terms of importance. For instance, we may define that mobility prediction information is three times more important than the relative size of the mobility group. In that case, the cell that compares mobility prediction with relative size of the mobility group will have a value of 3/1, and the opposite comparison the value of 1/3. This matrix, however, cannot be used directly. An eigenvector with the final weights has to be calculated following the procedure:

1) Convert fractions to decimals.
2) Square the resulting matrix.
3) Sum up the rows of the matrix and get a vector. Each of the rows of the vector must be divided by the sum of all its rows to normalize the values.
4) Repeat the previous steps until the resulting vector is not different from the previously obtained vector.

With the scores of each alternative for each criterion and the weights, the score of alternative i is given by:

$$S_i = \sum_{\substack{j=1 \\ \forall i \in [1,N]}}^{M} w_j r_{ij} \tag{1}$$

where:

S_i is the score of the i^{th} alternative;

r_{ij} is the normalized rating of the i^{th} alternative for the j^{th} criterion, which is calculated as $r_{ij} = x_{ij}/(\max_i x_{ij})$ for benefits and $r_{ij} = \frac{1}{x_{ij}}/(\max_i \frac{1}{x_{ij}})$ for costs;

x_{ij} is an element of the decision matrix, which represents the original value of the j^{th} criterion of the i^{th} alternative;

w_j is the weight of the j^{th} criterion;

M is the number of criteria;

N is the number of alternatives.

With the list of destination alternatives ranked and sorted in descending order by their score, hereafter just called "ranking", the decision about where to migrate content may be made based on the defined policies. For instance, the first three alternatives (destinations) of the ranking can be selected and content will be migrated to all of them. Or, depending on the problem and assuming that the score is normalized, alternatives with scores above 0.75 are to be selected.

After the decision about the destinations has been taken, the remaining decision of **what** content should be migrated still needs to be taken. This decision has to be made considering that currently the association between LTE users and NDN users is not known, and therefore content cannot be related to a particular moving user. Therefore, it takes the local popularity of content as key criterion [21] and considers the following steps. First, if the content object being considered is available nearby (1 hop distance) or already at the destination, it will not be migrated. Second, if it is not available, and if free space is available at the destination, content objects are just migrated until the cache is filled. Third, if no free space is available, both popular content at the origin and destination should be considered together and ranked to fill the destination cache with the content that will deliver the greatest benefit for all the users (existing and new ones). That problem can be modeled as a Knapsack problem, and be solved with Dynamic Programming [28]. However, it is a NP-complete problem and, even if a solution is found, it may take too long to calculate. Therefore, another simpler approach was followed. The following equation was considered to calculate the score of each content object k:

$$S_k = \begin{cases} p_k * \dfrac{n_{mgt}}{n_{mgt} + n_{dst}}, \text{if the } k^{th} \text{content object is} \\ \text{considered for migration.} \\ \\ p_k * \dfrac{n_{dst}}{n_{mgt} + n_{dst}}, \text{otherwise.} \end{cases} \tag{2}$$

where:

S_k is the score of the k^{th} content object;

p_k is the local popularity of the k^{th} content object;

n_{mgt} is the number of users migrating from origin to the selected destination;

n_{dst} is the number of users at the selected destination for content.

With the content objects ranked and sorted in descending order by their score, content is selected to fill the cache until its size threshold. When content is not already available locally, decisions are made towards deciding **how** to copy it from its nearest replica and **when** that operation should be performed. The first decision is based on a simple load balancing strategy, considering the available links' status information and giving priority to direct links, e.g. 3GPP-defined X2 interfaces between eNBs. The second decision is derived from the available time for migration (given by mobility prediction) and the existing schedule for other migrations using the same components. Based on the time it will take to copy the content subset and the deadline to have it copied, a slot is picked in

the migration schedule and the FMC Controller instructs the NDN router at the destination to fetch the content accordingly.

IV. Evaluation Experiments

In order to evaluate the proposed strategies, a set of experiments was defined and is described in detail in the next subsections.

A. Mobility Data Input

In order to evaluate the proposal described in the previous section, a realistic mobility trace was selected [29]. This trace includes data from one hundred human subjects over the course of nine months, and it was collected by MIT students using Nokia 6600 smart phones in the academic year of 2004/2005. Although the information collected includes call logs, Bluetooth devices in proximity, cell tower IDs, application usage, and phone status, for this evaluation only the information on mobility was considered: Object Identifier (OID), endtime, starttime, person_OID, celltower_OID. Therefore, it is possible to know at every given time to which cell a given person is connected. Such trace can thus be used to assess how the system behaves in realistic conditions, as if mobility was generated in any other way, it would probably create biased results and render the conclusions invalid for real-world scenarios.

Concerning mobility prediction, it is not the main focus of this work. Therefore, 15 minutes delta time predictions (15 minutes in the future) were generated at every 1 hour of simulation time according to the results obtained by mobility prediction works [30][31]. As concluded in the mentioned works, a 50% user movement randomness corresponds to an accuracy of about 50%. Thus, the generated predictions for these experiments had an accuracy following a normal distribution $N(0.5, 0.1)$.

B. Basic Setup

The setup for this evaluation is depicted in Fig. 2. It consists of the proposed architecture implemented in the ns3 simulator using its LTE module [32] together with ndnSIM 2.1 [33]. First, the simulator creates a Content Producer attached to a NDN Router, which is a node with NDN capabilities such as caching and forwarding. The latter is by itself attached using IP and 10 Gbps links to the EPC of the LTE module. Afterwards, a pair of eNB + NDN Router (including a NDN Content Store, i.e. cache of 2 GB stored in RAM) is created for each cell of the trace mobility file, attaching randomly positioned (within the cell's coverage) UEs + NDN Consumers to the LTE network according to the trace mobility inputs. These attachments are changed over the simulation time, thus emulating user mobility and triggering an handover using the X2 interface. That handover is managed by the MME, which is modified to feed information to the Mobility Predictor. The Mobility Predictor feeds mobility information, while NDN Routers provide the remaining relevant information (criteria) to the FMC Controller, which makes decisions and therefore instructs content to be copied between NDN Routers.

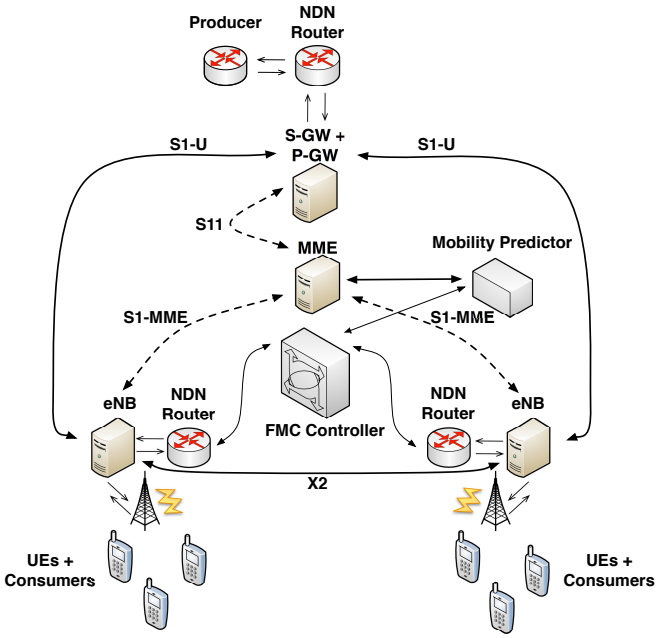

Fig. 2: Evaluation Setup

As for the simulation itself, it runs for 12 hours in a daytime period of the trace mobility file, when it is more likely for users to be active. Simulations were repeated 30 times using different day periods with at least 60 active users and considering the 100 most visited cells, using a Linux cluster to parallelize the work (http://www.ubelix.unibe.ch). Additionally, for parameters not mentioned here, the default values were used (e.g. LTE radio parameters).

C. Decision Criteria's Weights

Assuming that different weights for the criteria may return different results, four different weight sets were considered for evaluation. These are highlighted in the tables below, which each contain the AHP judgment matrix for all the criteria and the resulting eigenvector with weights calculated using the process described in Section III.

	M.P.	Diff	F.S.	G.S.
M.P.	1/1	2/1	4/1	3/1
N.C.	1/2	1/1	2/1	3/1
F.S.	1/4	1/2	1/1	1/2
G.S.	1/3	1/3	2/1	1/1

(a) Matrix

	Weight
M.P.	0.46124
N.C.	0.28450
F.S.	0.10633
G.S.	0.14793

(b) Vector

TABLE I: Weight Set #1

	M.P.	Diff	F.S.	G.S.	Cost
M.P.	1/1	2/1	4/1	3/1	3/1
N.C.	1/2	1/1	2/1	3/1	2/1
F.S.	1/4	1/2	1/1	1/2	1/1
G.S.	1/3	1/3	2/1	1/1	2/1
Cost	1/3	1/2	1/1	1/2	1/1

(a) Matrix

	Weight
M.P.	0.39778
N.C.	0.25232
F.S.	0.09725
G.S.	0.14897
Cost	0.10368

(b) Vector

TABLE II: Weight Set #1 with Cost

In Table I, the importance given to mobility prediction (M.P.) is higher than for any other criterion. At the same time, group size (G.S.) is considered more important than free space (F.S.) and cost is not considered.

	M.P.	Diff	F.S.	G.S.
M.P.	1/1	1/2	3/1	2/1
N.C.	2/1	1/1	3/1	4/1
F.S.	1/3	1/3	1/1	2/1
G.S.	1/2	1/4	1/2	1/1

(a) Matrix

	Weight
M.P.	0.28450
N.C.	0.46124
F.S.	0.14793
G.S.	0.10633

(b) Vector

TABLE III: Weight Set #2

	M.P.	Diff	F.S.	G.S.	Cost
M.P.	1/1	1/2	3/1	2/1	2/1
N.C.	2/1	1/1	3/1	4/1	3/1
F.S.	1/3	1/3	1/1	2/1	1/1
G.S.	1/2	1/4	1/2	1/1	2/1
Cost	1/2	1/3	1/1	1/2	1/1

(a) Matrix

	Weight
M.P.	0.31739
N.C.	0.36398
F.S.	0.11526
G.S.	0.10714
Cost	0.09623

(b) Vector

TABLE IV: Weight Set #2 with Cost

In Table II, everything is similar to Table I besides the fact that migration cost is now considered and the resulting weights are different.

In Table III, the greatest importance is given to the amount of non-intersecting content between the caches (N.C.) and, unlike in Table I, F.S. is considered more important than G.S. Also here, cost is not considered.

In Table IV, everything is similar to Table III besides the fact that migration cost is now considered and the resulting weights are different.

D. Content and Requests

The Content Producer consists of a file generator, which generates 100 000 files according to the defined scenario: either a YouTube scenario or a web server scenario. The first scenario intends to mimic video streaming traffic using conditions from the well-known YouTube video portal, which is the type of traffic that dominates Internet nowadays. The second scenario attempts to mimic traffic of users accessing modern Web 2.0 pages with plenty of multimedia content such as high-resolution images. As shown in Table V, it is assumed that content popularity of both of them follows a Zipf distribution [34]. For this setup, 20 popularity classes are taken into account. As several studies have shown [35][36] that most content objects are unpopular and only a few content objects are very popular, the number of content objects to be included in each popularity class is mapped to a Zipf distribution with $\alpha = 1$ and with inverted classes, i.e., most content objects are included in class 19 and fewest files in class 0.

Parameter	Web Server	YouTube
Requests	Every 5 seconds	
Request Popularity	Zipf distribution with $\alpha = 1$	$\alpha = 2$
File Distribution per Popularity Class	Zipf distribution, $\alpha = 1$ mapped to inverse classes	
File Sizes per Popularity Class	Gamma distribution, $\alpha = 1.8, \beta = 1200$ min. 50KB max. 50MB	Gamma distribution, $\alpha = 1.8, \beta = 5500$ min. 500KB max. 100MB

TABLE V: Evaluation Parameters

As also described in Table V, file sizes within each popularity class are different. Based on existing YouTube models [37], file size distribution for a YouTube scenario is set to a gamma distribution with $\alpha = 1.8$ and $\beta = 5500$. The file sizes for web server traffic are considerably smaller [38]. However, these file sizes have increased during the last years, and it is safe to assume that they keep increasing in the future with NDN. Transmitted NDN packets need to have a certain minimum size to be efficient, e.g., segment size of 4096 bytes or more, to avoid too large overhead for content headers including names and signatures. Therefore, it is assumed that for future NDN traffic, many small files may be aggregated to larger data packets or NDN would only be applied to large static files, e.g., pictures or embedded videos, and not small text files that may change frequently. Therefore, a gamma distribution with $\alpha = 1.8$ and $\beta = 1200$ was selected for the web server scenario.

As for the content requests to be performed by users (Consumers) during the simulation, a parameter of $\alpha = 1$ is considered realistic for web server traffic and $\alpha = 2$ is used for YouTube traffic.

E. Evaluation Metrics

Finally, five different metrics were evaluated to assess accuracy of the strategies, performance improvements for end users and potential savings for operators. The first is the position of an optimal solution (highest profit destination) in the score-sorted ranking of destination alternatives derived from the output of the AHP decision. The optimal solution is the location where the group of users was on which the requested volume of content recently migrated was the highest. If it is in the first positions (1, 2, 3, etc.) of the aforementioned ranking, it means that the decisions were good and will yield benefits for the users. The second metric is the number of cache hits, which enables the comparison of strategies and the benefit to be quantified in terms of end users perspective and possible network bandwidth savings. The third is the average download latency experienced by users, considering the best weight set from previous metrics evaluation. The fourth is the aggregated usage of bandwidth at the core interfaces (S1 and X2), evaluating the overhead caused by different strategies and how the load becomes distributed. Finally, the fifth is a comparison of timings for FMC in the different scenarios in order to evaluate if migrations are made on time when they are reactive (no predictions) or proactive (mobility predicted).

V. Evaluation Results

In the graphs below, results from the experiments defined in the previous section can be observed. To assess the first evaluation metric a comparison is made between the different weight sets, and a Cumulative Distribution Function (CDF) is generated for each ranking position. With the CDF, it is possible to obtain the cumulative percentage of times when the optimal solution was within the n first positions of the ranking. For example, one may assume that x percent of the times the optimal solution was at the first three positions of the ranking of destinations.

In Fig. 3, results show that weight set #1 has a higher percentage of optimal solutions at the first position of the ranking, meaning that selections were perfect in almost 60%

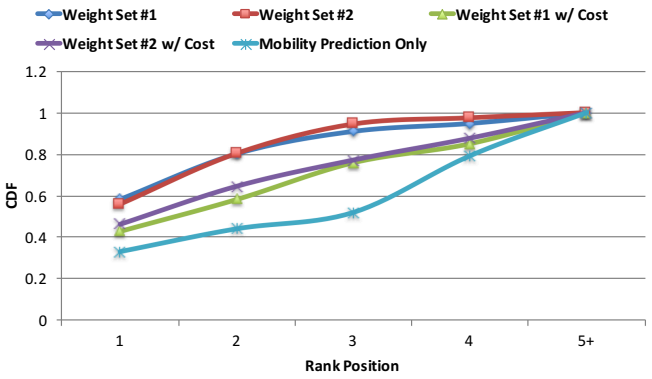

Fig. 3: Optimal Solution in Ranking

of the cases. However, the results of weight set #2 converge quicker to 100% of the cases, surpassing weight set #1 after (and including) rank position number 2. Overall, one may assume that weight set #2 selects better options than any other weight set, especially considering that the optimal solution was within the first three positions of the ranking in more than 90% of the cases. This can be easily explained by the accuracy of mobility prediction, which can vary immensely and does not account for the time users spend at the predicted locations. At the same time, giving priority to destinations where most of cached content is not the same as in the origin has a big inherent potential to be explored from the beginning.

When looking at the weight sets but considering cost, the trend is slightly different. Weight set #2 with cost outperforms weight set #1 with cost from the beginning, with the optimal solution being in the first position of the ranking almost 50% of the cases and in more than 80% of the cases the optimal solution being in the first 4 positions of the ranking. These results are according to what was expected, as cost limits the performance but considering it still delivers a good trade-off for both end users and network operators.

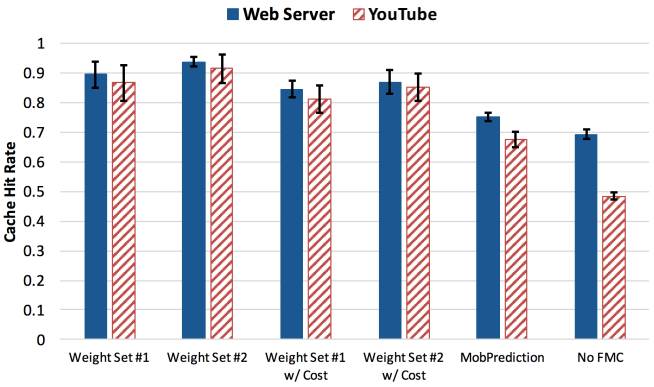

Fig. 4: Cache Hit Rates

As for the second evaluation metric, a comparison between our strategies with different weight sets, mobility prediction only [22] and no use of FMC (default strategy) is shown in Fig. 4, evaluating cache hits in both the YouTube and Web Server scenarios.

From the depicted results, one may observe that cache hit rates tend to be lower for the YouTube scenario because of bigger file sizes and a different Zipf distribution. However, in this particular case our FMC strategies show the biggest difference towards simple Mobility Prediction and No FMC. For instance, the cache hit rate using weight set #2 is up to 40% higher than with the default strategy without FMC, and over 20% higher than relying solely on Mobility Prediction.

As for the Web Server scenario, the benefit is not so high (up to 20% less). Such fact is explained by the characteristics of web server traffic, which has a lot of small objects that are easily cached even if the cache storage space is low. Therefore, users may find most of the content already distributed over the network, and migration strategies do not copy a large amount of content that can yield benefits. However, multimedia content now accounts for the most traffic in mobile networks [1], and we can easily conclude that FMC content migration strategies deliver their biggest performance for the biggest part of the traffic.

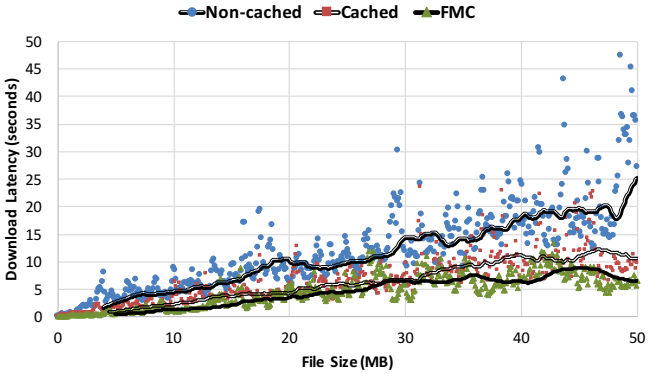

Fig. 5: Average Content Download Latency - Web Server

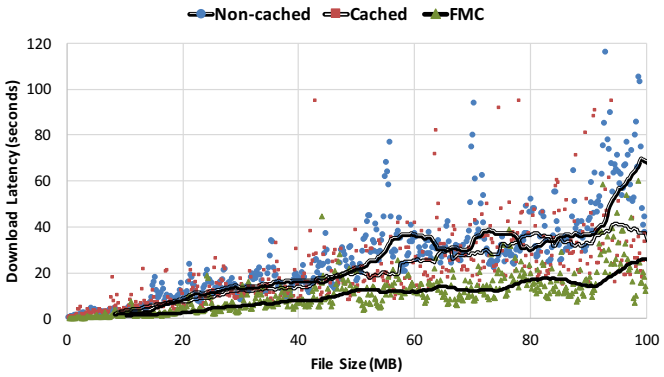

Fig. 6: Average Content Download Latency - YouTube

In Fig. 5 and Fig. 6, a comparison between the different strategies (no edge caching, edge caching and FMC) is presented for the Web Server and YouTube scenarios. Here, the FMC strategy is considered to be Weight Set #2 with cost, thus having the highest cache hit rates together with possible savings in core bandwidth. First, we may observe that data points present high variance, caused by the method they were obtained with. As the number of files is too big to represent, sampling was performed to include only 500 data points between the minimum and maximum file sizes. This sampling considers the sizes of all files generated in the multiple runs, and therefore has the influence of the different file sizes

themselves, network conditions, processing and others. Thus, to facilitate the understanding of the results, a trend line with a moving average of 50 data points is included. Results confirm in a end-to-end user perspective what was visible when comparing cache hit rates: improvements experienced by end users are considerable and caches are used more efficiently, i.e. cached content corresponds mostly to content that will actually be requested by users. This is true for both scenarios, and again we may easily see that improvement towards regular edge caching is much bigger when multimedia traffic is considered and content sizes tend to increase.

As for aggregated core bandwidth usage, Fig. 7 depicts a comparison for the different strategies (no edge caching, edge caching and FMC), in the two scenarios and also in different LTE core interfaces (S1-U and X2). First, we observe that, as expected, the aggregated usage of the S1-U interfaces is clearly reduced for both scenarios when caching at the edge. Second, we can also see that there is an overhead created by using FMC strategies when comparing to edge caching. This overhead becomes more clear over time, when caches start to be filled and more content is migrated, but it is compensated by the usage of the X2 interfaces between eNBs. This balances the load and eventually even adds more load to the X2 interface (prioritize), thus moving traffic away from the EPC and using available resources more efficiently. Finally, we conclude that there are differences between scenarios, especially because of the file sizes being bigger in YouTube traffic. This leads to a higher bandwidth consumption reduction with edge caching in the YouTube scenario, but also a slightly higher overhead for FMC strategies. Despite that, reductions in the usage of S1-U interfaces, and therefore in EPC, are still meaningful because of more traffic being offloaded using X2 interfaces.

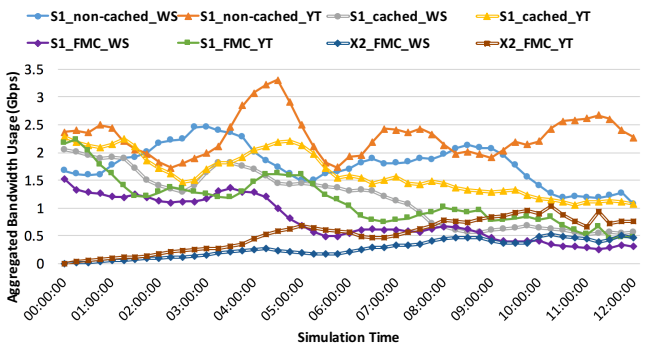

Fig. 7: Aggregated Core Bandwidth Usage

Finally, in Fig. 8 we plot the average execution times of different components for the FMC strategies (in both traffic scenarios) together with the average available time brought by both X2 handover procedures and mobility prediction. All values have a confidence interval of 95%, and we see that despite the accuracy of mobility prediction (about 50%), on average there is still plenty of available time for decisions and other cache operations (bear in mind that we are using a logarithmic scale). Using mobility prediction, in both scenarios the FMC components are able to execute within the available time frame. At the same time, if FMC operations are triggered

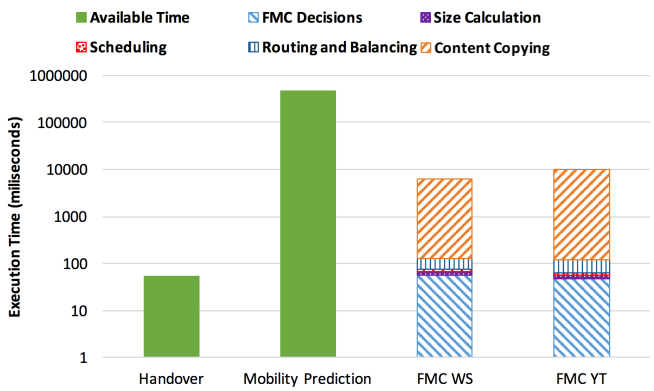

Fig. 8: Execution Times

reactively (no prediction), the handover delay is only enough for the decisions process. This means that the content transfer procedure will only start after the user is already at the new location, and in the worst case scenario it will have some initial cache misses while the content is still transferring. Overall, the impact of this behavior is not very high, as from previous figures we still observed very high cache hit ratios.

VI. Conclusions

In this paper, concepts and strategies for the migration of content within mobile networks were introduced, enabling multiple benefits both from user and network perspectives. As the users move to different locations, they still want to access content in which they are interested with a low latency and without delays or breaks, especially if dealing with multimedia content. From the network perspective, this can only be granted if caches exist at the edge of mobile networks and content kept in those caches (with limited resources) is the right content, i.e. popular content that local users are very interested in.

A number of proposals to handle this issue already exist, and were described thoroughly in Section II. However, some cannot be applied to content (only to services) or have other limitations, often assuming a very specific scenario or scope and not handling important issues or considering certain requirements. Therefore, we propose a broader approach to deal with content migration, handling decisions with multiple criteria and deciding multiple factors that will trigger content migration to a particular place of a given subset of content.

This proposal was evaluated in terms of performance, considering multiple weight values and different scenarios. When comparing to the case where default NDN caching strategies are used, clear benefits can be observed and quantified, leading to the conclusion that not only FMC enhanced caching strategies are the way to go when handling edge caches, but also that the architecture proposed in subsection III-A together with its decision mechanisms can achieve the goal of delivering content with lower latency to end users while efficiently using and saving well-valued network bandwidth.

Although the results can be considered as quite good, improvements can still be made. For instance, more hierarchical levels can be considered in the criteria for the decision where

to scale content and more advanced strategies can be used to decide which subset of content should be migrated. We envision that popularity may not be the only factor to decide which content to migrate due to is general nature, but also other factors that relate user to content should be considered in future work.

ACKNOWLEDGEMENT

The research leading to these results has received funding from the European Union's Seventh Framework Programme Mobile Cloud Networking project (FP7-ICT-318109).

REFERENCES

[1] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2014–2019," http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.pdf, Feb 2015.

[2] P. Demestichas, A. Georgakopoulos, D. Karvounas, K. Tsagkaris, V. Stavroulaki, J. Lu, C. Xiong, and J. Yao, "5g on the horizon: Key challenges for the radio-access network," *Vehicular Technology Magazine, IEEE*, vol. 8, no. 3, pp. 47–53, Sept 2013.

[3] "Suggestions on Potential Solutions to C-RAN by NGMN Alliance," The Next Generation Mobile Networks (NGMN) Alliance, Tech. Rep., Jan. 2013. [Online]. Available: http://www.ngmn.org/uploads/media/NGMN_CRAN_Suggestions_on_Potential_Solutions_to_CRAN.pdf

[4] B. Haberland, F. Derakhshan, H. Grob-Lipski, R. Klotsche, W. Rehm, P. Schefczik, and M. Soellner, "Radio Base Stations in the Cloud," *Bell Labs Technical Journal*, vol. 18, no. 1, pp. 129–152, 2013.

[5] V. Jacobson, D. K. Smetters, J. D. Thornton, M. F. Plass, N. H. Briggs, and R. L. Braynard, "Networking named content," in *Proceedings of the 5th International Conference on Emerging Networking Experiments and Technologies*, ser. CoNEXT '09. New York, NY, USA: ACM, 2009, pp. 1–12.

[6] S. K. Fayazbakhsh, Y. Lin, A. Tootoonchian, A. Ghodsi, T. Koponen, B. Maggs, K. Ng, V. Sekar, and S. Shenker, "Less pain, most of the gain: Incrementally deployable icn," in *Proceedings of the ACM SIGCOMM 2013 Conference on SIGCOMM*, ser. SIGCOMM '13. New York, NY, USA: ACM, 2013, pp. 147–158.

[7] D.-h. Kim, J.-h. Kim, Y.-s. Kim, H.-s. Yoon, and I. Yeom, "Mobility Support in Content Centric Networks," in *Proceedings of the Second Edition of the ICN Workshop on Information-centric Networking*, ser. ICN '12. New York, NY, USA: ACM, 2012, pp. 13–18.

[8] D. Smetters and V. Jacobson, "Securing network content," PARC, Tech. Rep., Oct. 2009. [Online]. Available: https://www.parc.com/content/attachments/TR-2009-01.pdf

[9] A. Gomes and T. Braun, "Feasibility of Information-Centric Networking Integration into LTE Mobile Networks," in *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, ser. SAC '15. ACM, April 2015, pp. 628–634.

[10] X. Wang, M. Chen, Z. Han, D. Wu, and T. Kwon, "Toss: Traffic offloading by social network service-based opportunistic sharing in mobile social networks," in *INFOCOM, 2014 Proceedings IEEE*, April 2014, pp. 2346–2354.

[11] M. P. Wittie, V. Pejovic, L. Deek, K. C. Almeroth, and B. Y. Zhao, "Exploiting locality of interest in online social networks," in *Proceedings of the 6th International COnference*, ser. Co-NEXT '10. New York, NY, USA: ACM, 2010, pp. 25:1–25:12. [Online]. Available: http://doi.acm.org/10.1145/1921168.1921201

[12] M. D. Choudhury, H. Sundaram, A. John, D. D. Seligmann, and A. Kelliher, ""birds of a feather": Does user homophily impact information diffusion in social media?" *CoRR*, vol. abs/1006.1702, 2010.

[13] H. Li and G. Ascheid, "Mobility prediction based on graphical model learning," in *Vehicular Technology Conference (VTC Fall), 2012 IEEE*, Sept 2012, pp. 1–5.

[14] S. Rajagopal, N. Srinivasan, R. Narayan, and X. Petit, "Gps based predictive resource allocation in cellular networks," in *Networks, 2002. ICON 2002. 10th IEEE International Conference on*, 2002, pp. 229–234.

[15] Y. Chon, H. Shin, E. Talipov, and H. Cha, "Evaluating mobility models for temporal prediction with high-granularity mobility data," in *Pervasive Computing and Communications (PerCom), 2012 IEEE International Conference on*, March 2012, pp. 206–212.

[16] A.-F. Antonescu, A. Gomes, P. Robinson, and T. Braun, "Sla-driven predictive orchestration for distributed cloud-based mobile services," in *Communications Workshops (ICC), 2013 IEEE International Conference on*, June 2013, pp. 738–743.

[17] T. Taleb and A. Ksentini, "Follow me cloud: interworking federated clouds and distributed mobile networks," *Network, IEEE*, vol. 27, no. 5, pp. 12–19, September 2013.

[18] ——, "An analytical model for follow me cloud," in *Global Communications Conference (GLOBECOM), 2013 IEEE*, Dec 2013, pp. 1291–1296.

[19] A. Ksentini, T. Taleb, and M. Chen, "A markov decision process-based service migration procedure for follow me cloud," in *Communications (ICC), 2014 IEEE International Conference on*, June 2014, pp. 1350–1354.

[20] H. Liu, Y. Sun, and M. S. Kim, "Provider-level content migration strategies in p2p-based media distribution networks," in *Consumer Communications and Networking Conference (CCNC), 2011 IEEE*, Jan 2011, pp. 337–341.

[21] C. Anastasiades, A. Gomes, R. Gadow, and T. Braun, "Persistent caching in information-centric networks," in *Local Computer Networks (LCN), 2015 IEEE 40th Conference on*, Oct 2015, pp. 64–72.

[22] X. Vasilakos, V. A. Siris, G. C. Polyzos, and M. Pomonis, "Proactive selective neighbor caching for enhancing mobility support in information-centric networks," in *Proceedings of the Second Edition of the ICN Workshop on Information-centric Networking*, ser. ICN '12. New York, NY, USA: ACM, 2012, pp. 61–66.

[23] "Named Data Networking (NDN) project," http://named-data.net/techreport/TR001ndn-proj.pdf, PARC, Tech. Rep., Oct. 2010.

[24] D.-L. Xu, "An introduction and survey of the evidential reasoning approach for multiple criteria decision analysis," *Annals of Operations Research*, vol. 195, no. 1, pp. 163–187, 2012.

[25] T. L. Saaty, *The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation*. New York, NY: Mcgraw-Hill, 1980.

[26] R. Saaty, "The analytic hierarchy process—what it is and how it is used," *Mathematical Modelling*, vol. 9, no. 3–5, pp. 161 – 176, 1987.

[27] T. L. Saaty, "How to make a decision: The analytic hierarchy process," *European Journal of Operational Research*, vol. 48, no. 1, pp. 9 – 26, 1990, decision making by the analytic hierarchy process: Theory and applications.

[28] R. Andonov, V. Poirriez, and S. Rajopadhye, "Unbounded knapsack problem: Dynamic programming revisited," *European Journal of Operational Research*, vol. 123, no. 2, pp. 394 – 407, 2000.

[29] N. Eagle and A. (Sandy) Pentland, "Reality mining: Sensing complex social systems," *Personal Ubiquitous Comput.*, vol. 10, no. 4, pp. 255–268, Mar. 2006.

[30] N. Samaan and A. Karmouch, "A mobility prediction architecture based on contextual knowledge and spatial conceptual maps," *Mobile Computing, IEEE Transactions on*, vol. 4, no. 6, pp. 537–551, Nov 2005.

[31] N. Amirrudin, S. Ariffin, N. Malik, and N. Ghazali, "User's mobility history-based mobility prediction in lte femtocells network," in *RF and Microwave Conference (RFM), 2013 IEEE International*, Dec 2013, pp. 105–110.

[32] "ns-3: LTE Module," https://www.nsnam.org/docs/models/html/lte.html, Sep. 2015.

[33] A. Afanasyev, S. Mastorakis, I. Moiseenko, and L. Zhang, "NS-3 based Named Data Networking (NDN) simulator," http://ndnsim.net, Sep. 2015.

[34] D. Rossi and G. Rossini, "Caching performance of content centric networks under multi-path routing (and more)," http://perso.telecom-paristech.fr/~drossi/paper/rossi11ccn-techrep1.pdf, Telecom ParisTech, Tech. Rep., 2011.

[35] T. Yu, C. Tian, H. Jiang, and W. Liu, "Measurements and analysis of an unconstrained user generated content system," in *Communications (ICC), 2011 IEEE International Conference on*, June 2011, pp. 1–5.

[36] "Half of youtube videos get fewer than 500 views," http://www.businessinsider.com/chart-of-the-day-youtube-videos-by-views-2009-5?IR=T, May 2009.

[37] A. Abhari and M. Soraya, "Workload generation for youtube," *Multimedia Tools Appl.*, vol. 46, no. 1, pp. 91–118, Jan. 2010.

[38] A. Williams, M. Arlitt, C. Williamson, and K. Barker, "Web workload characterization: Ten years later," in *Web Content Delivery*, ser. Web Information Systems Engineering and Internet Technologies Book Series, X. Tang, J. Xu, and S. T. Chanson, Eds. Springer US, 2005, vol. 2, pp. 3–21.