Caching in the Air: High Altitude Platform Stations for Urban Environments

Caching in the Air: High Altitude Platform Stations for Urban Environments

Greta Vallero, Daniela Renga, Michela Meo
Politecnico di Torino, Italy

***Abstract*—Due to the evolution in communications technologies and antennas, as well as advances in solar panel efficiency, High Altitude Platforms (HAPS) have been recently considered as a promising aerial network component, to support Radio Access Networks (RANs). Through their directional antenna they can activate beams and provide coverage to up to 1.5 km radius ground area. In this work, we consider a HAPS equipped with a Multi Access Edge Computing (MEC) server, which provides caching capabilities. The HAPS is used to off-load content requests. We analyse an urban environment scenario, as well as the effects of the simultaneous activation of beams in different areas. Results demonstrate that the HAPS is a suitable solution to bring additional capacity to the RAN and highlight that the provided performance strictly depends on the traffic demand profile of the covered portion of RAN.**

***Index Terms*—High Altitude Platform Station, Multi Access Edge Computing, Radio Access Network, 6G, 5G**

I. Introduction

Aerial networks are emerging as complementary infrastructure to the terrestrial Radio Access Network (RAN) that expands it in terms of network capacity and coverage to meet the needs of a wide range of services and applications [1]. As defined by the Third Generation Partnership Project (3GPP) in [2], these networks are part of the Vertical Heterogeneous Network (VHetNet), which is structured in three layers: satellites (space) network, aerial network, and terrestrial network. Since satellite communications suffer from high path-loss attenuation and significant propagation delays, the aerial network has been proposed and it relies on High Altitude Platforms (HAPS) and Unmanned Aerial Vehicles (UAVs). UAVs are typically drones and can operate between 100 and 400 m from the ground. Mounting a BS on them (UAV-BS) is a promising solution to bring connectivity where needed for short amounts of time, as in crowded scenarios or emergency situations [3]–[10]. The use of UAV-BSs is, however, limited to the scenarios in which the action is urgent and limited in time, since they generally fly for a time ranging from a few tens of minutes to a few hours, due to the scarce amount of on-board available energy.

HAPS' characteristics are in between satellite and UAVs and, for this reason, HAPS are receiving a lot of attention. In projects such as Google Loon, a HAPS is a network node, typically an airship or a balloon, which operates at an altitude between 20 and 50 km [11]. The HAPS has the potential of providing high data rates for a large coverage area with significantly lower latency than satellites, having also the economic advantage of lower development and deployment costs for the same coverage, with respect to satellite and terrestrial networks [12]. Lower altitudes than satellites means lower latency, making HAPS suitable for low-latency application [11]. This, combined with the fact that its altitude increases the probability of Line of Sight (LoS) links with users, results in low channel attenuation and, due to its almost stationary position, significant Doppler shift is avoided [1], [11]. HAPS operates above the clouds, where natural solar energy is abundant, and is typically large enough to host high capacity solar panels to generate large amounts of energy, which preserve its functionality for up to years [1]. As a result, the HAPS can stay afloat in the stratosphere for several months as a BS platform or as repeater, with the communication unit powered by solar panels [12]. HAPS are connected to users through access links, while they access the Core Network (CN), establishing Backhaul (BH) links between them and an Access point (AP). The HAPS uses Adaptive Antenna Array (AAA) or Active Electronically Scanned Array (AESA) antennas to dynamically form beams and, with each beam, provide coverage to a portion of a ground area, whose radius is up to 3 km [13]. In this way, it is possible to sectorizes a large service area to multiple cells.

While in most of the previous literature HAPSs have been considered as a means to reach uncovered remote areas, the activation of HAPS beams covering heterogeneous urban areas is usually neglected. In this paper, we investigate a Heterogeneous RAN scenario in which the HAPS carries both a BS and a MEC server to cache the most popular contents. The benefits of HAPS caching are investigated in terms of the Quality of Service (QoS) experienced by users, in case multiple areas are served simultaneously by the HAPS.

II. Methodology

A. Scenario

We consider a heterogeneous Long Term Evolution-Advanced (LTE-A) RAN, comprising many macro cell BSs. Each macro BS is supported by 6 micro cell BSs, whose radio coverage overlaps with their macro, meaning that small cell BSs are deployed to provide additional capacity. A HAPS is integrated in the RAN, and used as an aerial data centre, to off-load the content requests that are generated by the ground users. The HAPS and, if specified, ground BSs are equipped with MEC servers, to store some contents closer to the users. When a user requests a content, if that content is locally cached at the HAPS, it is directly delivered to the user. If this is not the case, it is retrieved in the cloud. The hardware technology of each cache is DRAM (Dynamic Random Access Memory) and the server updates its contents according to the Least

Frequently Used (LFU) cache algorithm, whose objective is storing the most popular contents. We assume that the file library is composed of 1000 files, with size equal to 100 Mbit. As in our previous works presented in [14], [15], the frequency and channel bandwidth of the considered radio technology, are 2.6 GHz and 5 MHz, respectively. Single Input Single Output (SISO) antennas for both the micro and macro cell BSs are considered. The link budget is as in [16]. The HAPS operate in the stratosphere, specifically at 20 km from the ground, as it is recommended by ITU-R in [13] to minimise the required propulsion power for keeping the HAPS stationary, given that at this altitude the wind speed is minimum. The HAPS uses an AESA antenna technology with a four facet panel system. It is composed of a tilted four facet panel configuration with four beams per facet, meaning that it provides up to 16 beams [12]. As a result, the HAPS dynamically sectorizes its large service area, whose diameter is larger than 20 km, to up to 16 small areas. Each of these small cells has a radius of 1.5 km and can be controlled individually, re-using the same frequency [12], [17]. We assume that each small beam covers a portion of ground RAN composed of a macro BS and its six micro cell BSs. As indicated in [13], [17], the HAPS operates at 27.9-28.2 GHz and, assuming that its spectrum efficiency is 4 bit/s/Hz, which is the target in [17], its total capacity is 1.2 Gbps, which means 75 Mbps/beam. The HAPS is supplied by an on-board PV panel system and an energy battery, large enough to satisfy the HAPS energy demand, without exceeding the maximum payload that the HAPS is able to carry.

B. Input data

To investigate the proposed scenario, we carry out trace-driven simulations in which real data about traffic are used. The data, which are provided by a large Italian mobile network operator, report the traffic demand volume, in bit, of 1420 BSs located in the city of Milan (Italy) and in a wide area around it, for two months in 2015, with granularity of 15 minutes. For each 15 minutes long time slot, the traffic volume in uplink and downlink is reported. The traffic traces are normalised, to shape the growth of the traffic demand since 2015, and aggregated, to achieve an hourly granularity. Hence, the peak of each traffic pattern is equal to the maximum capacity of each BS. For our work, we select eight portions of the city, each corresponding to the area covered by a single SMBS-HAPS beam. These areas are selected as samples of quite different traffic patterns ranging from residential to business, from touristic to campus areas. All together, the selected areas are representative of the various zones that coexist in an urban environment. In each of these portions of the RAN, we assume that one macro BS and 6 small cell BSs are present, so that the service area is covered by one macro cell which overlaps with 6 small cells.

C. Simulations

We simulate the scenario assuming that it operates for two months. In each simulation, the time is slotted in a 1 hour-long time interval. For each time slot, the traffic demand, as well as the content requests, are determined in each of the considered areas. In the areas which are covered by a HAPS beam, the content requests are offloaded to the HAPS. Finally, each cache is updated. Below the details of our simulations are given.

Generation of the traffic

In each time slot t, we assume that the traffic demand of each ground BS bs in each area is given by the sum of the uplink and downlink volume, respectively $T^{DL}_{bs,t}$ and $T^{UL}_{bs,t}$, in bit, provided by the data set described in section II-B. The number of content requests at time t, on BS bs, $N^{C}_{bs,t}$, is determined as $N^{C}_{bs,t} = \lfloor \frac{T^{DL}_{bs,t}}{S} \rfloor$, where $T^{DL}_{bs,t}$ is the downlink traffic volume at time t on BS bs, S is the size of each content and the $\lfloor \cdot \rfloor$ operator is needed to represent an integer number of requests. We consider a finite library $\mathcal{F}$=$\{1, 2, .., F\}$, composed of F content items, each with size S, in bits. Notice that this assumption can be easily removed. Each content has its popularity, which varies geographically. This means that each of the considered areas is characterised by a specific order of popularity of contents, which is modelled by the Probability Density Function (PDF) $P_{\mathcal{F},n}(f)$. As a consequence, it is possible that the probability that the content f is required in an area is different than in another one. Nevertheless, for each area, $\sum_{f=1}^{F} P_{\mathcal{F}}(f) = 1$. As in [18], [19], the popularity is described by a Zipf's distribution, characterised by the parameter α. This parameter affects the difference among contents in terms of popularity. In case the value chosen for α is large, the most popular contents are significantly more popular than the other contents, and by decreasing α the popularity of contents behaves more similarly to the uniform distribution.

Traffic Management

In each time slot t, once the content requests have been determined in each area, we verify if they are satisfied locally, by the ground MEC cache of that area. In particular, in each area, for each content request of each BS, we verify if the requested content is stored in the ground MEC cache server. If this is the case, a ground *hit* occurs and the required content is directly sent to the user. If this is not the case, a ground *miss* occurs and we verify if i) the HAPS has an active beam which is covering the considered area, ii) that beam can serve that request, i.e. that beam has enough available capacity to carry S, in bits. If both the conditions are verified, the content request is off-loaded to the HAPS. If the requested content is cached in the HAPS, a HAPS *hit* occurs and that content is directly transmitted to the user. In case of *miss*, i.e. the required content is not cached in the HAPS, the content is retrieved from the content provider, located in the cloud, reached by the CN, accessed through the BH links. When condition i) or ii) are not verified, the content request can not be off-loaded and is managed by the ground RAN, which retrieves the content in the cloud, passing by the BH links.

Cache Update

At the end of each time slot, the caches, both in the HAPS and in ground RAN, if present, are updated according to the Least Frequently Used (LFU) algorithm, so as to always cache the most popular contents. In the case of the cache in the HAPS,

the cache is updated considering the content requests on all the active beams.

D. KPIs

The following Key Performance Indicators (KPIs) are used to evaluate the performance. They are computed for each zone which is also covered by an active beam provided by the HAPS.

Average Access Traffic Handled by the HAPS

This is the amount of traffic of an area, in bits, carried by an active beam covering that area. It represents the traffic volume of the off-loaded content requests, generated in an area.

Average Access Traffic Handled by the Ground RAN

It accounts for the amount of traffic, in bits, carried by the ground RAN, covering an area, once content requests have been off-loaded to the HAPS.

Average BH Traffic Handled by the HAPS

In each area, we account for the volume of traffic, in bits, which passes through the BH network of the HAPS, i.e. the amount of traffic which the HAPS takes from the cloud to satisfy the off-loaded content requests.

Percentage of Requests Handled by the HAPS

It is the percentage of the content requests generated by users located in a given area that are off-loaded to the HAPS.

Percentage of Traffic Handled by the HAPS

It is the percentage of the total amount of traffic, in upload and download, from a given area that is off-loaded to the HAPS.

HAPS Miss probability

Given an area, we measure the miss probability of the off-loaded content requests, experienced by that area.

Total Miss probability

This is used to quantify the miss probability taking into consideration both the requests handled by the ground RAN of a zone and by the HAPS. It is given by the ratio between the number of misses of the requests handled by a given area and, if active, its covering HAPS beam, and the total number of content requests generated by that area.

III. Results

In this section, we discuss numerical results obtained by simulating the scenarios and methodologies presented in section II. Simulations start with empty caches, but the effects of this assumption is negligible because of the long operating time which we consider, equal to two months.

A. Effect of HAPS cache size

In the first part of our work, we assume that there is no MEC server in the ground RAN and the HAPS activates only two beams. We analyse the cases where the two active HAPS beams cover the Train Station and Rho Fiere (an exhibition area), as well as the Train Station and PoliMi (campus) zones. Given the heterogeneity of users who visit the Train Station and PoliMi areas, we use 0.56 as value of the α parameter of the Zipf's distribution used to model the content popularity, meaning that the contents have similar probability to be required. The Rho Fiere zone is characterised by low traffic demand with periods with significantly higher traffic demand than usual, which occur when fairs and exhibitions take place. We assume that these events attract users with

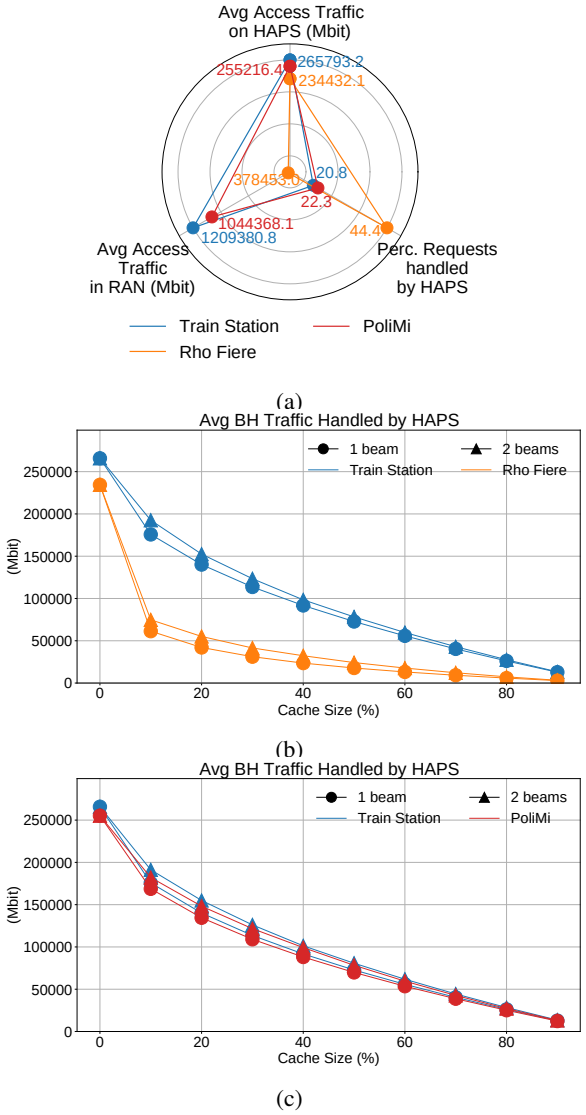

Fig. 1: KPIs with a single active beam in Train Station and Rho Fiere and PoliMi areas (a); Average BH Traffic handled by the HAPS in Train Station and Rho Fiere areas (b), in Train Station and PoliMi areas (c).

similar profiles and for this reason the α parameter can be set equals to 1.06, meaning that there are some contents which are quite more popular than the others.

Fig. 1a shows, for each area, the average access traffic handled by the ground RAN, the average access traffic handled by each beam of the HAPS, the percentage of content requests and traffic carried by the HAPS, in the Train Station, Rho Fiere and PoliMi areas. These KPIs depend on the characteristics of the traffic demand shape and are not affected by the capacity of the MEC server capacity, which is installed on the HAPS, nor by the number of simultaneous active beams. These figures indicate that the amount of the off-loaded traffic volume on the HAPS is similar in each zone, since content requests are served by the active covering beam, until its saturation. The HAPS serves more than 44%

TABLE I: Growth, in percentage, of the Avg BH Traffic Handled by HAPS for each beam, if two simultaneous beams are active and cover Train Station and Rho Fiere areas (on top), and Train Station and PoliMi areas (at the bottom).

Beams active on Rho Fiere and Train Station areas

(%)	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**
Train Station	9.5	8.8	8.6	7.3	7.5	7.1	6.5	5.4	3.4
Rho Fiere	22.0	31.0	33.2	37.0	36.1	35.1	31.9	25.3	13.7

Beams active on PoliMi and Train Station areas

(%)	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**
Train Station	8.8	10.6	11.0	10.5	11.0	10.8	9.9	8.4	5.4
PoliMi	7.9	10.2	11.5	12.7	12.1	11.2	10.2	8.0	5.9

of the content requests that are generated in Rho Fiere area, while in Train Station and PoliMi areas, no more than 21% and 22% of their content requests are carried by the HAPS, respectively. As a result, Train Station and PoliMi areas present a higher average access traffic volume in the ground RAN than in Rho Fiere.

In Figs. 1b and 1c, each curve corresponds to the average BH traffic handled by each active beam, in Mbit, when the MEC capacity on HAPS increases from 0, i.e. there is no cache on HAPS, to the case in which the cache stores 90% of the library. In these figures, the blue, orange and red curves show the values for the Train Station, Rho Fiere and PoliMi areas, respectively. In each figure, this is given when a single beam is active, see curves marked by circles, as well as when two beams are simultaneously active, see curves marked by triangles. In particular, in Fig. 1b the two simultaneous beams cover the Train Station and the Rho Fiere areas, while in Fig. 1c the HAPS provides at the same time the Train Station and the PoliMi areas with the service. First, observe how the growth of the size of the HAPS cache generates a reduction of the BH traffic volume, since more content can be stored locally. As largely demonstrated in the literature, this reduction strictly depends on the characteristics of the popularity, e.g., on the parameter α. Indeed, in the Rho Fiere area, which has α equal to 1.06, there is a small part of the library which is very popular. If this is the case, even a small cache drastically reduces the BH traffic handled by the HAPS. As can be seen in Fig. 1b, in case of a single HAPS active beam, which covers the Rho Fiere zone, the average BH traffic in the HAPS is reduced by up to 74%, if only 10% of the library is stored. In the Train Station and PoliMi areas the situation is different. The value of α is 0.56, meaning that the files have similar popularity and larger caches are needed to achieve a significant drop of the BH traffic carried through the HAPS. In this case, caching on the HAPS 10% of the library reduces its average BH traffic by only 34%, if a single beam is active.

B. Cache sharing between two HAPS cells

Figs. 1b and 1c highlight also the effect of the simultaneous activation of two beams that, as expected, increases the average BH traffic handled by the HAPS. Indeed, the HAPS cache is shared between the two areas served by the two active beams and a lower number of frequently used contents can be stored per each area. Table I reports, for each area and for each HAPS cache capacity, the growth, in percentage, of the average BH traffic handled by the HAPS, when a second beam is active and covers another area. The average BH HAPS traffic growth is significantly lower at the Train Station than in the Rho Fiere area. Indeed, while the growth in the latter is between 13% and 37%, in the former one it is never larger than 10%. The number of content requests in Rho Fiere is lower than in Train Station. In addition in Rho Fiere there is a very small fraction of contents which is very popular. Nevertheless, the remaining contents, which are the largest part of the library, are not so used to be cached on HAPS, when both the beams are active. For these reasons, the largest part of the HAPS cache stores contents which are popular in the Train Station area, making necessary the access to the cloud for the content requests generated in Rho Fiere. When the HAPS covers both the Train Station and the PoliMi area, see Table I, the average BH traffic grows with respect to the case which has a single active beam in a similar way in the two zones. As can be noticed, the growth of the Train Station is larger in this case than in case the second beam covers Rho Fiere. This is because PoliMi and Train Station off-load a similar amount of traffic, i.e. a similar number of content requests and these contents are required quite the same number of times. As a consequence, the HAPS cache stores about the same number of popular contents of these two zones.

From Table I, notice also that the lowest growth is observed for the smallest and the largest cache capacity. This is because with a small HAPS cache, the access to the cloud is needed for many content requests, as revealed by the high BH traffic handled by the HAPS in Figs. 1b and 1c. Activating a second beam, which means sharing the cache with another covered area, does not significantly increase the need to access the cloud to retrieve the requested contents. Similarly, when the cache stores up to 90% of the library, the cache is so large that it contains the most popular contents of both covered areas.

C. Integrating an on-ground MEC server

We now focus on the scenario in which the HAPS activates simultaneously the beams to bring connectivity to the Train Station and Rho Fiere areas. In addition to the aerial MEC server, we assume that an on-ground MEC server is installed at the Rho Fiere area and is accessible by each BS of that area. In Figs. 2a and 2c the miss probability from the HAPS and the miss probability from both the HAPS and the ground RAN are given, respectively, for the Train Station area, in blue, and the Rho Fiere area, in orange, increasing the ground MEC cache capacity from 0.0% to 50%. In these figures, each symbol used to mark each curve corresponds to a MEC cache

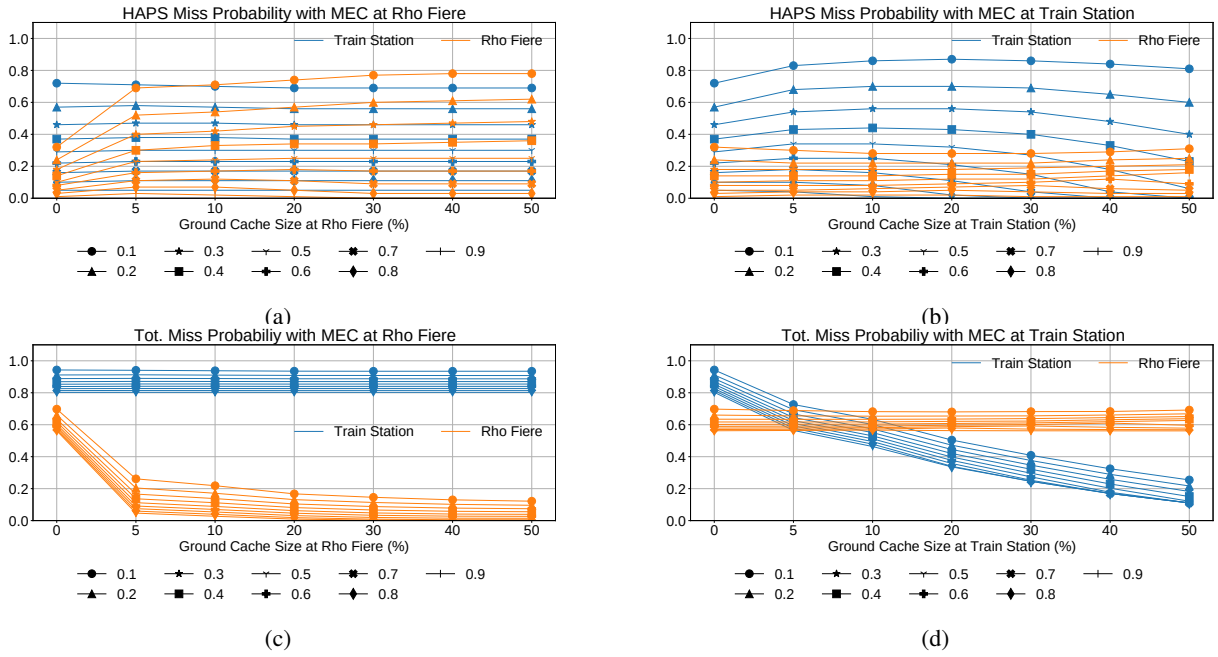

Fig. 2: Number of access to the cloud through the HAPS, if a MEC server in installed in the Rho Fiere ground RAN (a) and Train Station ground RAN (b); Total number of access to the cloud , if a MEC server in installed in the Rho Fiere ground RAN (c) and Train Station ground RAN (d).

capacity installed on the HAPS, from 10% to 90%. First, Fig. 2a confirms that Rho Fiere needs lower HAPS MEC cache capacity than Train Station to significantly reduce the experienced miss probability, when no ground cache is used (see the left part of the figure). This is because Rho Fiere has some contents which are quite more popular than the others, i.e. α is 1.06, while at Train Station the files have similar popularity, i.e. α equal to 0.56. Notice that small ground MEC cache in Rho Fiere, see Fig. 2a, causes the growth of the miss probability, which occurs on the HAPS, since fewer requests are off-loaded to it, making the contents of Rho Fiere less popular and, for this reason, not cached, further worsening the situation previously discussed. Meanwhile, from Fig. 2a it is possible to notice that the HAPS miss probability, experienced from the Train Station area, slightly decreases, while the Rho Fiere ground cache becomes larger, since more of its contents can be cached because they result more popular than those whose requests are forwarded from Rho Fiere. In addition, from Fig. 2c, we notice that for the Train Station the experienced miss probability results almost unchanged, while for the Rho Fiere area, it significantly drops because of the presence of the ground cache. Indeed, even a small cache capacity equal to 5% decreases it from 0.7, when there is no ground MEC to 0.26, if 10% of the library is stored on the HAPS.

In case a small ground MEC cache is installed in the Train Station, the HAPS miss probability tends to increase in this area, whereas a slight decrease is observed in Rho Fiere, as shown in Fig. 2b. Conversely, as the ground cache size grows larger, the HAPS miss probability experienced by the Train Station area decreases, especially under larger size of the aerial MEC server, whereas Rho area tends to register slightly higher values of HAPS miss probability. This is because more space in the aerial cache is occupied by a higher number of less popular contents that are requested by Train Station more frequently than popular contents requested by Rho Fiere, where contents are not requested enough to be cached. Finally, Fig. 2d highlights a symmetric behaviour in terms of total miss probability with respect to the configuration in Fig. 2c, although a more gradual descent in the Train Station area is observed as the size of the ground MEC server increases.

D. Cache sharing in multiple HAPS cell scenario

In the last part of our work, we let the number of active beams grow from one to eight. We start simulating a single active beam, which covers the Train Station area. Then, in each simulation, an additional beam is activated, covering a new area. For the content popularity distribution of central areas with heterogeneous traffic we assume that the α is 0.56. In two areas, corresponding to suburban exhibition areas, α is set to 1.06, since we assume that the public events which occur there attract users with a homogeneous profile. Each curve in Fig. 3 shows the average BH traffic handled by the HAPS for the Train Station, increasing the HAPS cache capacity, with the indicated number of active beams, from one to eight. As already discussed, the growth of the capacity of the MEC cache installed on the HAPS decreases the BH traffic and this drop depends on the characteristics of the traffic of the considered area, i.e., the α parameter. In addition, as previously discussed, increasing the number of active beams means rising the average BH traffic handled by the HAPS,

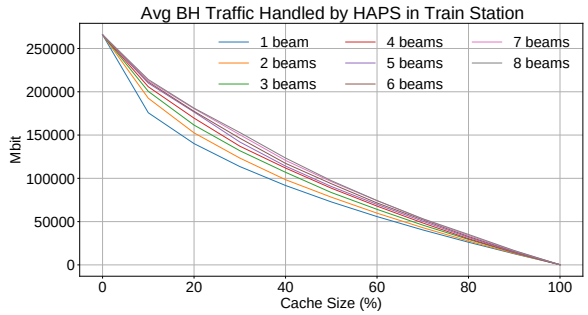

Fig. 3: Average BH Traffic handled by the HAPS for the Train Station areas, activating up to 8 beams.

since the cache is shared among the areas which are covered by one of its active beam, hence a larger fraction of quite popular contents from each traffic area cannot be stored in the HAPS MEC server. In the Train Station area, the average BH traffic handled by the HAPS accounts for 175 Gbit when the HAPS has a single active beam and is equipped with a cache server whose capacity is 10%. When there are 4 beams, 20% of the library has to be cached to obtain the same amount of average BH traffic. The growth of the average BH traffic handled by the HAPS in the Train Station area, with respect to the case where its beam is the only active one, is reported in Table II, for each HAPS cache capacity, increasing the number of active beams. From this table and from Fig. 3, we notice that the growth with the number of active beams is not constant. Indeed, when the number of active beams is small, adding a new one means significantly reducing the cache space of each area and this reduction becomes lower as the number of active beams increases. Table II highlights again what previously mentioned: the highest growth is obtained for very small and for very large values of the capacity, i.e. 10% and 90%.

IV. Conclusion

HAPS has been considered a valid support for the aerial networks, providing additional capacity to portions of a ground RAN, through directional antennas which can be dynamically oriented. In this work, we employ a HAPS equipped with a MEC server, providing caching capabilities, to off-load content requests generated in the ground RAN. The analysis reveals that a HAPS can be an effective solution to support the RAN, especially when the covered areas have common interests, and the sets of popular contents are similar. Nevertheless, if the coverage is brought simultaneously to heterogeneous areas, performance can be improved through the installation of ground MEC servers in those areas with fewer requests with local popularity.

TABLE II: Growth of the Avg BH Traffic Handled by HAPS for the Train Station areas, activating up to 8 beams.

(%)	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**
2 beams	9.6	8.9	8.6	7.3	7.5	7.1	6.5	5.4	3.4
3 beams	14.1	15.3	16.1	16.6	14.9	15.0	13.8	12.3	10.4
4 beams	17.1	21.0	20.7	22.0	21.8	21.7	20.0	15.6	12.7
5 beams	19.5	26.1	25.0	24.4	24.5	25.2	24.9	21.3	14.7
6 beams	20.3	26.6	28.8	27.7	29.0	28.2	29.0	25.4	21.4
7 beams	21.1	28.9	32.4	31.6	32.3	33.0	31.1	30.3	26.6
8 beams	21.9	29.2	32.6	34.6	33.8	33.2	32.1	34.4	29.4

References

[1] G. K. Kurt *et al.*, "A vision and framework for the high altitude platform station (haps) networks of the future," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 2, pp. 729–779, 2021.

[2] 3GPP, "Study on new radio (nr) to support non-terrestrial networks," 2019.

[3] G. Castellanos *et al.*, "Performance evaluation of direct-link backhaul for uav-aided emergency networks," *Sensors*, vol. 19, no. 15, p. 3342, 2019.

[4] ——, "Evaluation of flying caching servers in uav-bs based realistic environment," *Vehicular Communications*, vol. 32, p. 100390, 2021.

[5] R. Shakeri *et al.*, "Design challenges of multi-UAV systems in cyber-physical applications: A comprehensive survey and future directions," vol. 21, no. 4, pp. 3340–3385, 2019, conference Name: IEEE Communications Surveys Tutorials.

[6] M. Mozaffari *et al.*, "Beyond 5g with UAVs: Foundations of a 3d wireless cellular network," 2018. [Online]. Available: http://arxiv.org/abs/1805.06532

[7] B. Alzahrani *et al.*, "Uav assistance paradigm: State-of-the-art in applications and challenges," *Journal of Network and Computer Applications*, vol. 166, p. 102706, 2020.

[8] C. Qiu *et al.*, "Multiple uav-mounted base station placement and user association with joint fronthaul and backhaul optimization," *IEEE Transactions on Communications*, vol. 68, no. 9, pp. 5864–5877, 2020.

[9] ——, "Backhaul-aware trajectory optimization of fixed-wing uav-mounted base station for continuous available wireless service," *IEEE Access*, vol. 8, pp. 60 940–60 950, 2020.

[10] M.-J. Youssef *et al.*, "Full-duplex and backhaul-constrained uav-enabled networks using noma," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 9, pp. 9667–9681, 2020.

[11] M. S. Alam *et al.*, "High altitude platform station based super macro base station constellations," *IEEE Communications Magazine*, vol. 59, no. 1, pp. 103–109, 2021.

[12] G. Amitabha *et al.* Haps: Connect the unconnected. [Online]. Available: https://d1p0gxnqcu0lvz.cloudfront.net/documents/Nokia_Overview_of_High_Altitude_Platform_Stations_HAPS_White_Paper_EN.pdf

[13] ITU, "Technical and operational characteristics for the fixed service using high altitude platform stations in the bands 27.5-28.35 ghz and 31-31.3 ghz," ITU-Recommendation, Tech. Rep., May 2002.

[14] G. Vallero *et al.*, "Base station switching and edge caching optimisation in high energy-efficiency wireless access network," *Computer Networks*, vol. 192, p. 108100, 2021.

[15] ——, "Caching at the edge in high energy-efficient wireless access networks," in *ICC 2020-2020 IEEE International Conference on Communications (ICC)*. IEEE, 2020, pp. 1–7.

[16] M. Deruyck *et al.*, "Power consumption model for macrocell and microcell base stations," *Transactions on Emerging Telecommunications Technologies*, vol. 25, no. 3, pp. 320–333, 2014.

[17] ITU, "Spectrum needs of high-altitude platform stations broadband links operating in the fixed service," ITU-Recommendation, Tech. Rep., 2018.

[18] M. Chen, Y. Qian, Y. Hao, Y. Li, and J. Song, "Data-driven computing and caching in 5g networks: Architecture and delay analysis," *IEEE Wireless Communications*, vol. 25, no. 1, pp. 70–75, 2018.

[19] Z. Luo *et al.*, "Energy-efficient caching for mobile edge computing in 5g networks," *Applied sciences*, vol. 7, no. 6, p. 557, 2017.