Delay-based Core Network Placement in Self-Deployable Mobile Networks

Zhiyi Zhang, Razvan Stanica, Fabrice Valois

HAL Id: hal-03183703

https://inria.hal.science/hal-03183703

Submitted on 28 Mar 2021

Delay-based Core Network Placement in Self-Deployable Mobile Networks

Zhiyi Zhang*, Razvan Stanica*, Fabrice Valois*

*Univ Lyon, INSA Lyon, Inria, CITI, F-69621 Villeurbanne, France

{firstname.lastname}@insa-lyon.fr

Abstract—Self-deployable mobile networks represent a new type of cellular networks, that can be rapidly deployed, easily installed, and operated on demand, anywhere, anytime. They can provide network services when a classical cellular network fails, is not suitable, or does not exist. Using network virtualization techniques, core network and base station functions can be co-located together into a single equipment. This brings the network functions closer to the user, but the placement of the local core network has a significant impact on the network performance, mainly in terms of capacity and delay. In this work, we are focused on the delay minimization from BSs to the local core network. We propose a heuristic for the local core network placement, with the objective of reducing the delay. We show that the delay obtained by our solution decreases significantly compared to a strategy that places the local core network in order to maximize the capacity. We also show that considering a capacity-based placement strategy leads to infinite delay in some cases.

I. Introduction

Classical cellular networks are a result of careful planning and deployment strategies. Their fixed and hierarchical architecture is based on a clear physical separation between the radio access network (RAN) and the core network (CN), with an over-provisioned backhaul between them. If the network environment changes, a planned and time-consuming intervention is needed to adapt the network configuration. Therefore, classical cellular networks are not able to cope with dynamic changes and are difficult to use in some situations, like in the case of natural disasters. In these situations, mobile networks with self-deployment and self-configuration capabilities, which need only minor human intervention, are required. A possible architecture for self-deployable networks [1] is depicted in Fig. 1.

With Network Function Virtualization (NFV) and Software Defined Networking (SDN) playing a more and more important role in cellular networks, starting with 5G, core network functions can be provided by less complex and generic equipment. With such technologies, the network elements can be hosted in virtual environments to provide a higher degree of flexibility and adaptability, while network slicing protocols will direct the traffic to the best service functions, according to the specific use case. Therefore, in self-deployable mobile networks, there is no physical split between RAN and CN equipment: the CN functions could be virtualized and co-located with the base station (BS), forming a Local CN.
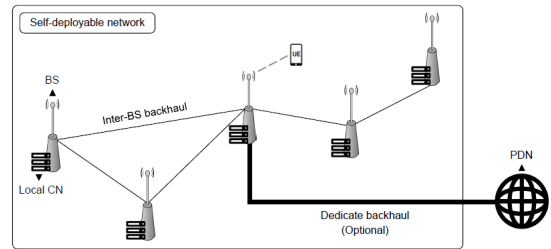
Fig. 1. Architecture of a self-deployable mobile network connected to an external Packet Data Network (PDN) [1].

A Local CN plays the same role as a core network in a classical cellular network, connecting users to the service network and implementing functionalities such as authentication or mobility management. Moreover, the equipment integrating the BS and the Local CN can be potentially carried by terrestrial robots or by unmanned aerial vehicles (UAV), giving the network infrastructure mobility capabilities. Therefore, the integration of the Local CN and the BS becomes an interesting research problem, with some studies already tackling the issue. The Local CN can be co-located with one single BS and shared by all the other BSs in the RAN, as proposed in [1], or one Local CN entity can be integrated at each BS, as proposed in [2]. In these previous works, research questions are focused on the capacity of the wireless links. However, delay is also a key important metric in mobile networks, especially for time-sensitive applications such as first responder interventions or live streaming. Previous works do not consider this metric, despite its significant impact. Indeed, the more data are transmitted, the higher the delay, which can cause numerous issues for time-sensitive applications.

In this paper, we propose a delay-based local CN placement in self-deployable mobile networks. Our objective is to find the best placement of the Local CN among the available BSs, in order to minimize the maximum delay in the backhaul network. This practically minimizes the worst case delay that a user can observe in a self-deployable network.

The rest of the paper is structured as below. Section II discusses some related research works on the topic of self-deployable networks. Section III introduces the system model and the delay problem formulation. Finally, Section IV shows the simulation results we obtained and Section V concludes the paper.

II. RELATED WORK

A first step towards the convergence between the radio access network and the core network is the proposal of the integrated access and backhaul (IAB) in the 3GPP Release 16 [3]. The idea is to share a part of the wireless spectrum between the RAN and the backhaul, in order to reduce the deployment costs of ultradense 5G mmWave networks.

The performance evaluation of IAB is investigated in [4]. The authors highlight the capacity of IAB to maintain the network performance (in terms of throughput and delay) while using less optical fiber to connect the BSs to the CN. The capital expenditure of the mobile operators is hence reduced. In their study, the RAN and the CN remain split between two separate entities. In our case, with self-deployable mobile networks, taking benefit of the CN co-located with the BS, the network performance should be increased. However, as we will show it, the location of the CN impacts significantly the network performance.

The impact of the placement of CN virtual functions on the operator costs is already demonstrated [5]. The optimal placement of VNFs in the core network has been well studied, both across federated clouds [6] and in sliced environments [7]. However, these studies consider the placement of CN functions in a classical architecture, with practically infinite backhaul. Instead, we consider a self-deployable architecture, where RAN and CN functions are co-located and carried by terrestrial or aerial vehicles.

The feasibility of mobile BSs seems well established. For example, ABSORB [8] is a cellular architecture consisting of BS transported by terrestrial vehicles, which can connect to an optical fibre backhaul network at different points in an urban area. The BSs change their position, in order to adapt to the traffic demand. Similarly, SkyCore [2] proposes to embed a BS and a Local CN in UAVs, interconnected by a wireless backhaul. Simulating the concept of mobile BS in the Milan urban area using realistic traffic patterns brings a gain of 120% in terms of throughput compared to a classical architecture [9]. Moreover, a self-depolyable approach has significant interest in the case of post-disaster scenarios [10] and it can even enable community owned and operated cellular networks in rural areas [11]. In the case of UAV-carried self-deployable networks, an important problem is the physical placement of the drones, with previous works focusing on coverage [12] and user mobility [13] issues.

Our study is complementary to those cited above, in the sense that we consider the placement of the physical BSs is given, and we tackle the problem of placing the Local CN. In this sense, SkyCore [2] considers the coexistence of multiple Local CNs, one per mobile BS. However, this leads to strange settings from a mobility management point of view, since users moving outside their main BS would practically be in a roaming situation, highly increasing the overhead. Therefore, using one single Local CN seems a better approach in self-deployable networks [1]. In [1], a centralized approach is used to compute the optimal placement of this unique Local CN. The objective of the authors in [1] is to maximize the traffic received by the Local CN from all the BSs in the network. For this, they introduce a new centrality metric, named flow centrality, which, when maximized, gives the optimal placement of the Local CN. However, this previous study completely ignores delay-related issues. In our work, we move one step forward, starting from the observation that, when the amount of data transmitted on the backhaul is close to the backhaul capacity, the transmission delay tends to infinity [14]. Therefore, we investigate the trade-off between capacity and delay for the Local CN placement.
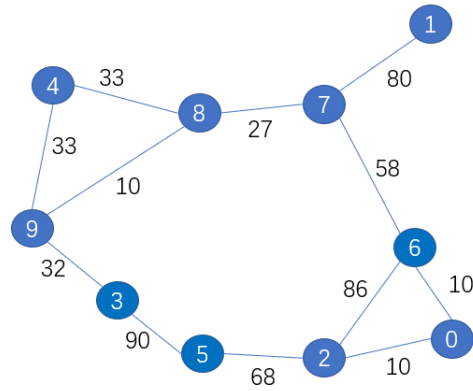

Fig. 2. A random geometric graph topology, with random link capacities $c(u,v) \in (0; 100]$.

III. SYSTEM MODEL

In self-deployable mobile networks, there is no traditional CN. All of the BSs should have a connection to the Local CN, which is co-located with one of these BSs. All the data traffic from BSs should be sent towards the equipment hosting the Local CN. We assume that all the BSs send the same amount of traffic towards the Local CN, and this traffic can be routed over different paths to reach the Local CN.

As in Fig. 2, we suppose that $G(N, \varepsilon)$ is an undirected graph modeling a mobile network, with $|N| = n$ nodes ($n = 10$ in the example in Fig. 2). Each node in the graph represents a BS and the edges connecting two BSs represent the wireless backhaul (e.g. a mmWave connectivity between BSs [15]). In this paper, based on insight from [1], we consider that there is only one Local CN in the network, and the destination BS hosting this Local CN is denoted by d.

Let $O, D \subseteq N$ be the set of sources and destinations, respectively: $O = N \setminus \{d\}$, and $D = \{d\}$. The inter-BS links have a limited capacity, represented by the maximum amount of traffic that can pass through the respective edge. We denote the capacity of the edge $(u, v) \in \varepsilon$ by $c(u, v)$: it means that a link is not able to manage more than $c(u, v)$ data traffic per time unit. However, the links are not necessarily used at their full capacity. Therefore, we denote the flow actually carried by an edge as $f(u, v)$. The total traffic (in data units per time unit) transmitted from a source node $v \in O$ to the destination d is denoted by $z(v, d)$.

To model the average packet delay in the network, we use the Kleinrock independence approximation [14]. Each inter-BS link is considered as a $M/M/1$ queue. We assume that the traffic generated by each source follows a Poisson process, the packet lengths are exponentially distributed and independent of each other. For $u \in O$, $v \in N$, let λ_{uv} be the Poisson arrival rate on the edge (u,v). For the service rate of the edge, denoted by μ_{uv}, we consider that it is related to the link capacity, i.e. $\mu_{uv} = c(u,v)$. Finally, t_{uv} is the processing and propagation delay on the given edge.

The average packet delay on an edge consists of three parts: the average waiting time in the queueing system, the average transmission time, and the processing and propagation delay. Therefore, the delay on the edge (u,v) is given by:

$$T_{uv} = \frac{\lambda_{uv}}{\mu_{uv}(\mu_{uv} - \lambda_{uv})} + \frac{1}{\mu_{uv}} + t_{uv} \qquad (1)$$

Let $p(s)$ be the path from the source node s to the destination node d. Let $\alpha_{s,uv}$ be the proportion of traffic generated by s carried by the link (u,v). In this case, the average packet delay from s to d, denoted by T_s, is given by:

$$T_s = \sum_{(u,v)\in p(s)} \alpha_{s,uv} T_{uv} \qquad (2)$$

The theoretical variation of the delay with respect to the traffic demand on a link with a given finite capacity, given by the model above, is shown in Fig. 3. The exact values on the two axes depend on several parameters, such as the link capacity or the propagation time. However, the important phenomenon to notice in the figure is that the delay increases as the flow on the link increases, and the higher the demand, the faster the delay grows. When the traffic demand on the link approaches its capacity, the delay will tend to infinity. Therefore, solutions focused on maximizing the traffic in the network [1] are not appropriate if delay is an objective. Indeed, if the focus is on maximizing the traffic, the network will be saturated, meaning that packets will encounter long delays.
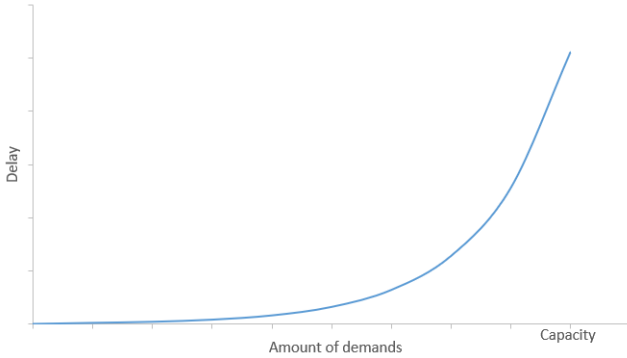

Fig. 3. Delay as a function of the amount of traffic demand on a given link.

A. Optimization Problem

In our model, all the nodes $s \in O$ send their data to the node d, where the Local CN is hosted. Among the average delays T_s

Node	0	1	2	3	4	5	6	7	8	9
Flow Centrality	2	8	18	12	7	12	17	10	7	8

TABLE I
FLOW CENTRALITY FOR EACH NODE IN OUR EXAMPLE.

observed by the nodes $\forall s \in O$, we denote the longest delay, produced by source node $\bar{s}$, as $T_{\bar{s}}$. Our objective is therefore to minimize $T_{\bar{s}}$, in order to contol the maximum delay in the network. We use as input the network topology $G(N, \varepsilon)$ and the amount of traffic generated at each source node $z(v,d)$. The optimal solution will give us the best placement of the Local CN d, as well as the route information $p(s)$ for each source node s to reach the destination d.

We can formulate the optimization problem as follows:

$$\min_{\bar{s}\in O} T_{\bar{s}} \qquad (3)$$

$$f(u,v) \leq c(u,v), \forall (u,v) \in \varepsilon \qquad (4)$$

$$\sum_{u\in O} f(u,v) + z(v,d) = \sum_{w\in N} f(v,w), \forall v \in O \qquad (5)$$

$$\sum_{v\in O} f(v,d) = (n-1) \times z(v,d) \qquad (6)$$

$$\begin{aligned} T_s = \sum_{(u,v)\in p(s)} \alpha_{s,uv} &\left(\frac{f(u,v)}{c(u,v)(c(u,v) - f(u,v))} \right. \\ &+ \left. \frac{1}{c(u,v)} + t_{uv} \right), \forall s \in O \end{aligned} \qquad (7)$$

$$T_{\bar{s}} \geq T_s \qquad (8)$$

The objective function is given in Eq. 3: the minimization of $T_{\bar{s}}$. The constraint given in Eq. 4 ensures that the flow passing through an edge does not surpass the capacity of the edge in question. The constraint in Eq. 5 is a simple flow conservation rule: the sum of the amount of traffic entering the node v plus the amount of traffic generated by node v should be equal to the amount of traffic exiting the node. The constraint in Eq. 6 indicates that the total amount of traffic received at the Local CN node d should be equal to the total amount of traffic generated by all source nodes $s \in O$. Eq. 7 allows us to calculate the average delay, based on Eq. 1 and Eq. 2. Finally, the constraint in Eq. 8 signals that $T_{\bar{s}}$ is the maximum value among all T_s.

B. Heuristic Solution

Unfortunately, the optimization problem formulated in Section III-A is not tractable by classical solvers, even using column generation or signomial geometric programming. This problem is not linear: the route to the destination for all s is not determined because, when the route used by a given node changes, the routes of all other nodes can be impacted.

Therefore, since this problem could not be handled directly by a solver, we decided to turn towards an heuristic approach
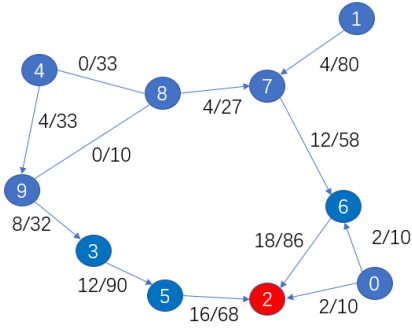
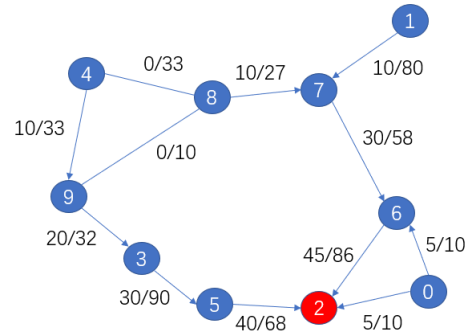Fig. 4. Route with minimum delay, demand=4, Local CN co-located with base station 2.

Fig. 5. Route with minimum delay, demand=10, Local CN co-located with base station 2.
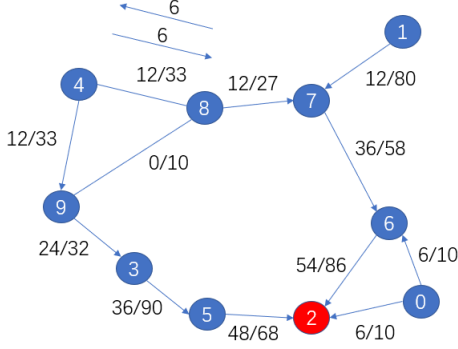

Fig. 6. Route with minimum delay, demand=12, Local CN co-located with base station 2.
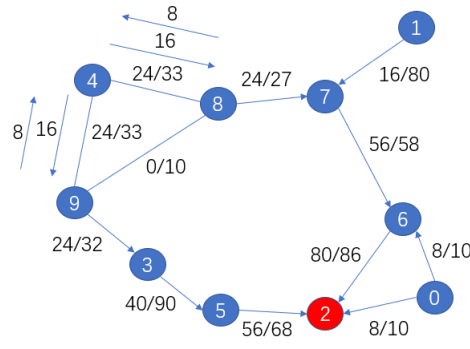

Fig. 7. Route with minimum delay, demand=16, Local CN co-located with base station 2.

to find a suitable solution to the problem. Considering a network topology and the limited capacities of the backhaul links, we compute the flow centrality [1] for each node, and we denote the group of nodes with the highest flow centrality metric as the Local CN candidates. Then, for each potential Local CN candidate, we vary the traffic demand at the source nodes and we try the different routing paths configurations. For each configuration, we compute the maximum end-to-end delay. Finally, the output of this heuristic is the configuration providing the lowest maximum end-to-end delay.

Note that, because we consider only a subset of locations for the Local CN and a subset of routing paths, there is no optimality guarantee in our solution. However, since we consider as possible candidates all the nodes providing a high flow centrality, our heuristic reaches an interesting trade-off between capacity maximization and delay minimization.

For example, considering the network topology shown in Fig. 2, we compute the flow centrality for each node, as shown in Table I. The maximum flow centrality is 18 for node 2, meaning that each node can send at most 18 units of traffic to a Local CN hosted by node 2. Thus, in our simulation, we vary the traffic demand at each source s using the values: $z(s,d) = \{4, 6, \ldots, 16\}$.

Next, using the nodes with the highest flow centrality in Fig. 2 as candidate destinations d, we list all the routes from s to d, $\forall s \in O$. Potentially, the flow sent from s to d is distributed over several paths. We select a limited number of routing

configurations. More precisely, in this example, we consider, for every source s the possibility of transmitting 100% of the traffic on all existing paths, as well as the possibility of splitting the traffic 50%-50% over all the existing pairs of paths. We then select the best routing configuration from the list, such as $T_{\bar{s}}$ is minimized for a given demand. Fig. 4 depicts the routing configuration obtained with our heuristic under the assumption of an homogeneous traffic demand equal to 4 traffic units: in this case, the node 2 is the Local CN.

At a close inspection of Fig. 4, in the previous example, the routing configuration looks like a shortest path protocol, except for the case of node 0. As a matter of fact, node 0 has only two exit links: $(0, 2)$, which goes directly to the Local CN, and $(0, 6)$ which represent a longer path, with an intermediate step. The capacity of both these links is limited to 10 traffic units, a relatively reduced capacity with respect to the other links in the network. Although the demand is only equal to 4 traffic units, this would represent 40% of a link capacity. As shown in Fig. 3, when the demand approaches the link capacity, the delay tends to increase significantly. This is why our results indicate, non-intuitively, that it is better to split the traffic demand over the two links. Practically, when the capacity of backhaul links is small, it is better to split the traffic over several paths, in order to reduce the load per link and, with it, the end-to-end delay.

IV. Performance Evaluation

A. Delay variation for a given placement of the Local CN

Considering the system model described in Section III, we assume t_{uv} = 0.01 (time units), such that the processing and propagation delay is not the dominant delay component. Also, Eq. 1 is not defined when $\mu_{uv} = \lambda_{uv}$, so we arbitrarily set $T_{uv} = 2$ time units in this case. For the network topology given in Fig. 2, we compute all $T_{\bar{s}}$ as a function of the demand when the Local CN is located with BS 2. We compare our results to the solution obtained in [1], where the Local CN placement is only based on the capacity maximization (note that, in this case, the optimal Local CN placement is also BS 2).

As expected, the maximum delay $T_{\bar{s}}$ increases exponentially as the traffic demand approaches the network capacity, as shown in Fig. 8. This result is consistent with the theoretical result expressed previously in Fig. 3. Considering the solution provided in [1], which is only focused on the capacity optimization without consideration for the delay, the maximum delay is equal to 4.21 time units, for a traffic capacity equal to 18 traffic units. However, if we accept to consume a little bit less capacity (16 units instead of 18), the delay decreases by more than 80%, while only 10% of the throughput is lost.
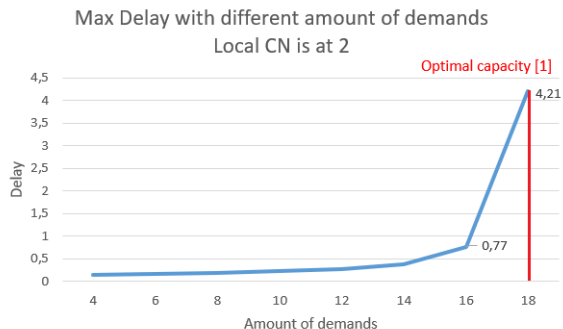

Fig. 8. $T_{\bar{s}}$ with different amount of demands.

For an even lighter demand, equal to 4 traffic units, the worst $T_{\bar{s}}$ value is obtained for node 4, as shown in Fig. 9. This is because node 4 has the longest path to the Local CN, which results in the longest delay. The second longest delay is for node 0, even if the distance from 0 to 2 is one of the shortest in the topology. Indeed, the traffic sent by the node 0 follows links $((0,6)$ and $(0,2))$ quite limited in capacity (see Fig. 4), thus creating an important delay.

For a demand in the range [6, 14], the distribution of T_s looks quite similar. For example, Fig. 10 shows the results for a demand equal to 10 traffic units. As observed previously, the path length and the use of limited backhaul links lead to an important end-to-end delay, typically for nodes 4 and 0. As the demand increases, the delay on capacity-constrained links increases exponentially. In this case, a load-balancing routing approach can help keep the end-to-end delay reasonably low. This can be observed in Fig. 5 and Fig. 6, depicting the routing paths and the traffic distribution for a traffic demand of 10 and 12 units, respectively. When the demand evolves from 10

to 12, nodes 4 and 8 reduce their end-to-end delay through a load-balancing routing. In this sense, Fig. 11 shows that, when the demand is equal to 16 traffic units, close to the maximal value of 18, the maximum end-to-end delay $T_{\bar{s}}$ is observed by nodes 4, 8 and 9. The corresponding paths are shown in Fig. 7, where we notice that load balancing is used.

B. Delay comparison with different Local CN placements

In the case of a centralized Local CN placement [1], each node in the network topology can have a different flow centrality metric, as it is shown in Table I. Because node 2 has the maximum flow centrality (18 units), it is chosen as the location of the Local CN. In addition, node 6 has a flow centrality metric of 17, and nodes 3 and 5 reach a metric of 12 traffic units, all significant values. In the following, we study the minimum value of $T_{\bar{s}}$ for different traffic demands, when nodes 2, 3, 5 and 6 are candidates for the Local CN placement.

As shown in Fig. 12, if the Local CN is co-located with node 2 or node 6, the delay is minimal, and there is no significant difference between the two cases, which is consistent with the very close flow centrality metric values. If the Local CN is located at nodes 3 or 5, the delay is significantly more important (20% extra delay for node 5 and 60% for node 3). Clearly, to consider only the node providing the maximum throughput, as in [1], reduces the degree of freedom in the Local CN placement problem. Considering both throughput and delay, two nodes can be selected as a potential Local CN in this example.

C. Backhaul load with different Local CN placement strategies

If different Local CN locations impact the end-to-end delay, they also impact the traffic load of backhaul links, due to the routing paths and the traffic distribution. We investigate the backhaul load, especially for the links which receive a higher load: the links which are directly connected to the Local CN.

In Fig. 13, we compare the load of the three links (0,2), (5,2) and (6,2), when the Local CN is co-located with base station 2, and considering two scenarios. In the first one, we consider the maximum flow centrality metric, focusing on throughput maximization, whereas in the second one, we consider our heuristic which also considers the delay. With our proposition, the backhaul load is reduced by 18%, which is really significant, especially if we consider that two of the links were at full capacity, a very dangerous situation from a networking point of view.

V. Concluding Remarks

In traditional cellular networks, all the BS configurations need to be carefully planned before the deployment. If the network fails, an interruption of communications occurs. Self-deployable mobile networks can play a role in this case, because they can automatically be deployed by using mobile robots, providing a temporary or persistent connectivity for end users. If the capacity remains the main performance indicator, delay is also critical, especially for real-time and
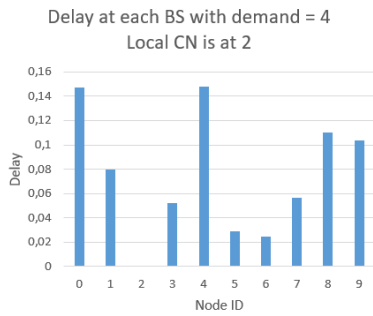
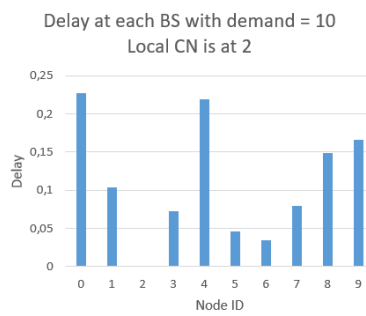Fig. 9. T_s distribution for paths with minimum delay, demand=4.

Fig. 10. T_s distribution for paths with minimum delay, demand=10.
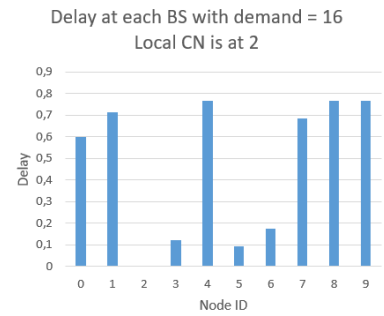

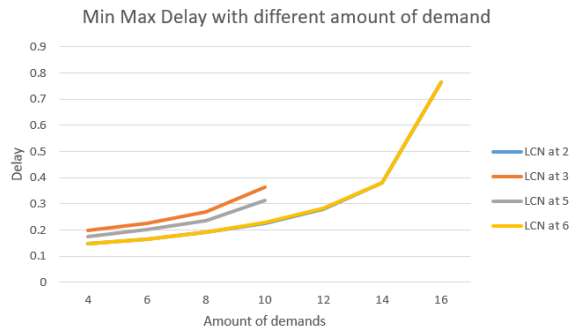Fig. 11. T_s distribution for paths with minimum delay, demand=16.

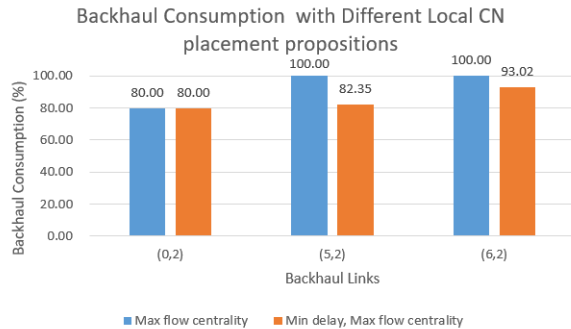Fig. 12. Min Max Delay with different amount of demand.

Fig. 13. Backhaul consumption Comparison with Different Local CN placement propositions.

time-sensitive applications. In this paper, we propose a delay-based Local CN placement in self-deployable mobile networks. More precisely, we deal with a trade-off between capacity maximization and delay minimization.

As it is well-known in queuing theory, the delay increases exponentially with the load. Unfortunately, we show that, when we consider only the capacity maximization, some links are saturated, leading to an explosion of the end-to-end delay. We also show that, the lower the link capacity, more often such situations occur. Therefore, we exploit path diversity to balance the traffic, especially when the link capacity is low. Finally, in our results, we also highlight that a small reduction

in terms of capacity can help reaching a reasonable delay.

In future work, we look forward to explore more complex topologies and to consider more routing paths combinations.

REFERENCES

[1] J. Oueis, V. Conan, D. Lavaux, H. Rivano, R. Stanica, and F. Valois, "Core network function placement in self-deployable mobile networks," *Computer Communications*, vol. 133, pp. 12–23, Jan. 2019.

[2] M. Moradi, K. Sundaresan, E. Chai, S. Rangarajan, and Z. M. Mao, "SkyCore: moving core to the edge for untethered and reliable UAV-based LTE networks," in *ACM MobiCom*, Oct. 2018.

[3] 3GPP, "NR - study on integrated access and backhaul - Rel. 16," *TR 38.874*, 2018.

[4] M. Polese, M. Giordani, T. Zugno, A. Roy, S. Goyal, D. Castor, and M. Zorzi, "Integrated access and backhaul in 5G mmWave networks: potential and challenges," *IEEE Communications Magazine*, vol. 58, pp. 62–68, Mar. 2020.

[5] F. Z. Yousaf, P. Loureiro, F. Zdarsky, T. Taleb, and M. Liebsch, "Cost analysis of initial deployment strategies for virtualized mobile core network functions," *IEEE Communications Magazine*, vol. 53, pp. 60–66, Dec. 2015.

[6] A. Baumgartner, V. S. Reddy, and T. Bauschert, "Mobile core network virtualization: a model for combined virtual core network function placement and topology optimization," in *IEEE NetSoft*, Apr. 2015.

[7] S. Agarwal, F. Malandrino, C. F. Chiasserini, and S. De, "VNF placement and resource allocation for the support of vertical services in 5G networks," *IEEE/ACM Transactions on Networking*, vol. 27, pp. 433–446, Feb. 2019.

[8] Y. Nakayama, T. Tsutsumi, K. Maruta, and K. Sezaki, "ABSORB: autonomous base station with optical reflex backhaul to adapt to fluctuating demand," in *IEEE Infocom*, May 2017.

[9] F. Mohammadnia, C. Vitale, M. Fiore, V. Mancuso, and M. Ajmone Marsan, "Mobile small cells for adaptive RAN densification: preliminary throughput results," in *IEEE WCNC*, Apr. 2019.

[10] M. Y. Selim and A. E. Kamal, "Post-disaster 4G/5G network rehabilitation using drones: solving battery and backhaul issues," in *IEEE Globecom Workshops*, Dec. 2018.

[11] S. Sevilla, M. Johnson, P. Kosakanchit, J. Liang, and K. Heimerl, "Experiences: design, implementation, and deployment of CoLTE, a community LTE solution," in *ACM MobiCom*, Oct. 2019.

[12] J. Qin, Z. Wei, C. Qiu, and Z. Feng, "Edge-prior placement algorithm for UAV-mounted base stations," in *IEEE WCNC*, Apr. 2019.

[13] C. H. Liu, X. Ma, X. Gao, and J. Tang, "Distributed energy-efficient multi-UAV navigation for long-term communication coverage by deep reinforcement learning," *IEEE Transactions on Mobile Computing*, vol. 19, pp. 1274–1285, Jun. 2020.

[14] D. P. Bertsekas and R. G. Gallager, *Delay Models in Data Networks*, ch. 3, pp. 149–269. Prentice Hall, 2 ed., 1992.

[15] R. Li and P. Patras, "Max-min fair resource allocation in millimeter-wave backhauls," *IEEE Transactions on Mobile Computing*, May 2019.