Edge Nodes Infrastructure Placement Parameters for 5G Networks

1st Alejandro Santoyo-González,
Department of Network Engineering
Universitat Politècnica de Catalunya (UPC)
Barcelona, Spain
alejandro.santoyo@entel.upc.edu

2nd Cristina Cervelló-Pastor
Department of Network Engineering
Universitat Politècnica de Catalunya (UPC)
Barcelona, Spain
cristina@entel.upc.edu

***Abstract*—5G scenarios entail stringent requirements such as ultra-low latency and ultra-high reliability. Consequently, bringing computing, storage and networking resources to the edge of the network has become a key element for 5G deployment. However, capital and operational expenditures need to be carefully taken into consideration to achieve cost-effectiveness in this scenario. With edge nodes geographically distributed, efficiency is directly linked to the placement and capacity planning of such nodes. To develop an efficient strategy to deploy Edge Computing infrastructure for 5G services, the first step is to define thorough site selection criteria. Therefore, this paper proposes a set of parameters tailored to the evaluation and optimization of the edge nodes location selection process under a merged 5G and Edge Computing ecosystem.**

***Index Terms*—Edge Computing, 5G, NFV, Fog Computing, MEC, Cloudlet, Placement Optimization**

I. Introduction

In recent years, Edge Computing (EC) has become a key topic as core solution to stringent 5G technical requirements (e.g. ultra-low latency). Consequently, several models and implementations have been developed under this technological umbrella. Cloudlet Computing (CC), Fog Computing (FC) and Mobile Edge Computing (MEC) are the main examples. In spite of their differences, all these cutting-edge proposals have in common a geographically distributed set of nodes bringing computing, storage and networking resources to the network edge. This raises critical concerns regarding capital and operational expenditures (CAPEX and OPEX), deployment strategies and placement parameters. Presumably, thousands of Edge Nodes (ENs) are to be distributed within any large city in upcoming 5G scenarios. Reducing CAPEX and OPEX in this context is challenging as typical infrastructure placement costs are mixed with strong latency restrictions and ultra-dense deployment architectures. The problem starts by a current lack of adequate parameters to cost-effectively evaluate potential and unforeseen EN locations. Furthermore, the highly dynamic nature of the service demand in future use cases requires a deep evolution of the measuring and modeling approaches for both location and capacity planning processes.

This work has been supported by the Ministerio de Economía y Competitividad of the Spanish Government under the project TEC2016-76795-C6-1-R and AEI/FEDER, UE.

Current placement methods, including data center placement and facility location problems (FLPs) do not consider a sufficiently tailored and comprehensive parameter set for accurate EN site evaluation because: **a)** most FLPs formulations accounts for a cost reduction considering a quite limited set of factors (e.g. transportation distances), **b)** ENs have not yet been clearly defined (neither functionally nor physically), but their energy consumption patterns and small size when compared to traditional data centers, makes unfeasible to use traditional data center placement strategies, as those are mainly planned based on expected power consumption, **c)** most cache server distribution studies limit their scope to find the best logical deployment spot, excluding physical elements from the analysis such as location-dependent costs, **d)** non-technical site restrictions are mainly overlooked (e.g. social/political factors are not analyzed), **e)** efficiently placing ENs should merge a discrete and continuous search technique to assess both potential pre-identified locations and unforeseen but feasible sites within the territory of interest.

This paper aims to overcome the aforementioned issues. Our contributions comprise a suitable set of parameters proposal for cost-effective EN placement mechanisms. Furthermore, we provide key guidelines and insights for the EN deployment under a convergent ecosystems merging 5G, Network Function Virtualization (NFV) and EC. In summary, Section II shows relevant related research to the presented problem, Section III presents an overview of the EN placement problem (ENPP) and the proposed set of placement parameters. Finally, Section V recapitulates the findings.

II. Related Work

The groundwork for defining an EN placement parameter set for 5G environments, starts by the extensive research available on closely related topics such as: small to large-sized data center, Base Station (BS) and server placement strategies, FLPs solutions, 5G use cases and demands, mobile network characterization models, 5G enablers and EC inter-operation and deployment models.

Upcoming 5G networking extends the typical throughput-based design criteria (used in prior mobile network planning) to a 3-D performance metric cube based on throughput, number of links and delay [1]. Diving into this metric cube, the total delay of a packet transmission in a cellular network can be attributed to the RAN, fronthaul, backhaul, core network, and data center or external server [2].

ETSI standard in [3] describes the performance metrics to be improved by deploying a service on a MEC environment. These performance metrics are categorized as: (1) functional metrics and, (2) non-functional metrics. The former comprises latency, energy efficiency, throughput, goodput, packet loss, jitter, and QoS. The latter includes but is not limited to service availability, reliability and service processing load. Although the proposed metrics are referred to MEC systems, accurate extrapolation can be made for the ENPP. The research in [4] dives into current standardization and advances in Multi-access Edge Computing while pointing out the importance of parameters such as latency, bandwidth and costs as optimization metrics.

Data center placement has been extensively studied. Two of the few publicly available papers studying data center placement optimization are [5] and [6]. On the former, the authors formulate the problem as a linear programming model seeking to minimize the costs of the entire data centers network. As a particularity, they assumed as inputs the maximum number of servers to deploy and the user per server ratio. This article eases the comprehension of most of the key physical aspects of service infrastructure placement such as energy consumption, build and land costs, among others. In addition, from [6] the main goal is to obtain a placement baseline for all the components of a fog network based on micro data centers and a long-reach passive optical network.

Thorough research about FLPs is available in [7]. In general, FLPs consider demands and distance (transportation costs) as main parameters for optimization purposes, although capacitated FLPs also take into account limited facility capacities as a problem constraint. On the other hand, mobile network planning and specifically BS placement have been previously addressed [8]. On this topic, the main planning concerns comprise interference and energy consumption in the radio layer, coverage, capacity, traffic patterns characterization and demand geo-distribution.

Under the MEC paradigm umbrella, Enhanced Small Cells (SCeNBs) and other concepts and platforms like the proposed in [9], significantly differ in their deployment location considerations. While some solutions (Small Cell Clouds and Mobile Cell Clouds) assume to place the computation capacities within the RAN sites, others maintain the approach of a farther away location of the resources at centralized data centers but introducing new components and inter-working procedures to ensure better performance.

Very few articles are available about EC service infrastructure placement. Yannuzzi et al. [10] analyze the placement of fog nodes in the specific context of a city like Barcelona. The pursued goal is to cope with the requirements of smart cities by deploying fog nodes to satisfy broadly distributed use case scenarios such as event-based video and traffic management. In [11] the OPEX is reduced by minimizing the number of required gateways while satisfying predefined QoS demands. From [12], a framework for the edge servers placement assumes that service users are somehow clustered and edge sites are proposed as near to the optimal locations for each cluster as possible. Capacity provisioning is addressed through an ILP formulation based on the number of users. Meanwhile, the service demand variation is taken into consideration by analyzing demand pattern input data. Such baseline information includes details about traffic aggregation points (TAPs) geo-distribution and their underlying demand requirement values.

III. Edge Nodes Placement

There is no current consensus among the operators and the scientific community on what an EN should be. Nevertheless, it is widely accepted that it should be NFV-capable. This has a direct impact on the placement strategy as it implies some particular considerations regarding location-dependent costs and energy consumption, just to mention two examples. Overall, a thorough study of convergent enabling technologies, current and future traffic and service trends, scalability planning, top-level and low-level architectures and inter-component synergistic are some of the critical aspects to be considered to solve the ENPP. A study of the main elements linked to the EN site selection process is presented in the following subsections. Table I shows a summary of the parameters proposal.

A. Latency

Latency has been widely studied in the context of mobile networks and 5G use cases [2], [13]. However, under the ENPP, latency control entails certain particularities and complexities that must be addressed to satisfy ultra-low delay requirements.

The "edge" definition is still unclear for EC implementations (with the exception of MEC where edge servers and RAN nodes are co-located) and thus, the first challenge is to define the delay values that can be reduced through the EN placement optimization. The top level in Fig. 1 depicts a "traditional" overview of the communication channel from a mobile user to a service hosted in an EN. The total unidirectional transmission time of a 5G system depends on [14]:

- L_{radio}: the radio layer packet delay, it occurs between the base station and user equipment (it includes the Transmission Time Interval which must be less than $1\ ms$, propagation delay, signal processing time at the receiver, and re-transmission time due to packet errors).
- $L_{Fronthaul}$: the delay between the base station front-end and the centralized Baseband Unit (BBU), if applicable.
- $L_{Backhaul}$: also called backhaul delay, it is the time taken to traverse the core network entities and gateways.
- L_{Core}: core network processing time.
- $L_{Transport}$: communication delay between the core network and the cloud/edge service host.

What is more, authors in [14] claim that the user plane latency over-the-air should be set to a maximum of $0.5\ ms$ on average. For EC, the latency optimization is to be carried

TABLE I
EN PLACEMENT PARAMETERS

Parameter	Description
Latency	Round trip latency between any TG-EN pair
Throughput	TG throughput demand served by a given EN
Location-dependent costs	Costs directly linked to the site itself (e.g. land acquisition, build costs)
Location restrictions	Non-technical restrictions and additional site-related considerations
Reliability	Site and coverage reliability based on the location characteristics and its covered TGs reliability demands
Service Area Type	Coverage area classification (e.g. urban and rural)

out from the traffic aggregation points such as the BBU in a cloud-RAN deployment. Therefore, the EN site selection optimization could improve the RAN-to-EN delay (calculated as $L_{Fronthaul} + L_{Backhaul} + L_{Transport}$) for mobile networking and the TAP-to-EN delay for other network architectures [1]. As a result, considering the evolution of the mobile network core towards 5G presented in [15], namely a fully merged NFV/SDN architecture, three main scenarios could be expected (see Scopes A, B and C in Fig. 1).

a) Scope A: The service hosts (e.g. cache servers) or the virtualized mobile core components are deployed in a distributed manner within the ENs set. Within this scope, from a functional point of view, the User-EN communication could even occur without involving any core network entities, thus mostly excluding current L_{Core} delays. This way, when selecting a site to deploy an EN and considering a management and orchestration framework able to efficiently route traffic to the nearest core component through SDN-based mechanisms, $L_{Transport}$ becomes negligible. Consequently, the delay suitable for optimization through the EN placement strategy accounts for the sum of $L_{Fronthaul}$ and $L_{Backhaul}$. This means that only those EN locations where $(L_{Fronthaul} + L_{Backhaul}) \leq L_{max}$ can be selected as EN sites (where L_{max} is the maximum delay allowed between any User-EN pair including the related processing delays).

b) Scope B: The edge infrastructure comprises the service hosts, core network components and the BBUs presumably as Virtual Network Functions (VNFs). Therefore, both $L_{Backhaul}$ and $L_{Transport}$ are minimized and only $L_{Fronthaul}$ can be optimized through an efficient and accurate EN location selection strategy. Similarly to Scope A, the service path in Scope B could exclude any mobile core entities from being involved (as the User-EN communication may not necessarily involve the network core), and the core processing delays could be avoided or reduced. Additionally, RAN processing delays under this scenario may be minimized through the hosting of BBUs within the edge infrastructure.

c) Scope C: Each RAN site is allocated computing, storage and networking capacities (it is upgraded to EN). As a pure co-location strategy is followed, no optimization

[1]From this point on, TAPs are assumed to include RAN nodes

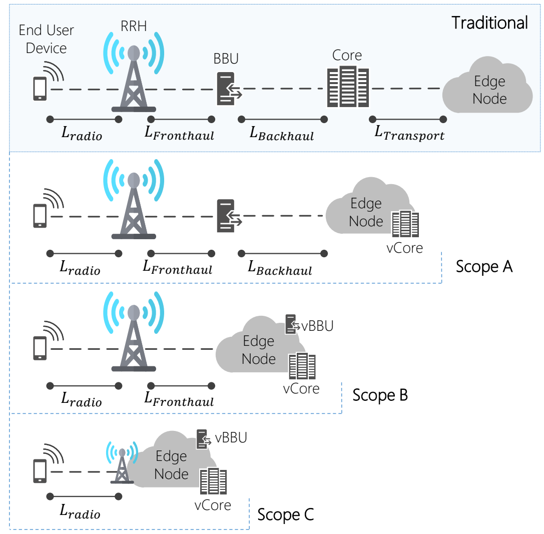

Fig. 1. Edge Node deployment scenarios for latency optimization.

is achieved by solving the ENPP (e.g. more cost-effective potential sites, such as Central Offices, are not considered and unforeseen locations are overlooked).

At first glance, **Scope C** offers the best deployment solution as it maximizes the latency reduction. However, this deployment scheme is not feasible due to scalability and cost related issues. As the number of TAPs will grow exponentially in 5G networks due to ultra-dense networking, the CAPEX and OPEX for upgrading each aggregation site to EN make this approach unfeasible. Furthermore, non-mobile service requests would not benefit from such placing strategy and thus 5G use case demands could not be entirely satisfied. In the case of **Scope A**, the limitation comes from not optimizing $L_{Backhaul}$ which is critical in order to achieve round-trip latency values under $1\ ms$. Moreover, the scenario in **Scope A** does not follow current 5G deployment advances and trends where the core components and the virtual BBUs coexist under the virtualized edge infrastructure. Taking these elements into account, **Scope B** can be assumed to be the most cost-effective EN deployment scheme, regarding latency optimization.

Overall, when placing an EN under 5G latency constraints, the maximum allowable delay between any aggregation point and its serving EN is of critical importance. Assuming a $L_{max} = 1\ ms$ threshold for delay-sensitive use cases and $L_{radio} = 0.5\ ms$, the ENPP solution should ensure $L_{max} = L_{radio} + L_{Fronthaul} + L_{processing} \leq 0.5\ ms$, considering $L_{processing}$ to be input data. This latency constraint is quite challenging [16]. However, a joint effort mixing radio-communication advances, EC placement optimization, service and management layer efficiency and other cutting-edge technologies, may result in turning such latency value a reality. In fact, recent research [17] has set a promising starting point towards achieving this goal.

Further latency complexities are introduced by a federation of edge and centralized cloud platforms, where under a hierarchical cloud arrangement, the operations with a local scope are handled by edge platforms while broader decisions are centralized. Such architecture can be seen as an extension of the traditional cloud, allowing flexibility in service deployment and mobility, by enabling an elastic combination of different resources across separate platforms for particular application types. This deployment requires an orchestration system to manage, control and configure the corresponding services across the set of cloud platforms.

B. Throughput

The capacity of an EN[2] directly depends on the traffic density. Such metric is tightly coupled with the 5G strict bandwidth requirements and expected ultra-dense device geo-distribution. In order to effectively consider traffic density for site selection purposes, the **network throughput** should be considered as a placement parameter.

One of the key ENPP trade-offs rises from the interrelation between throughput and EN capacity sizing. In principle, allocating as much demand as possible to each EN is desirable. Following this approach, commonly used base station placement strategies and tessellation mechanisms become suitable solutions [8], [18]. However, latency restrictions could then lead to unmet requirements and performance issues and location-dependent costs could be overlooked. Furthermore, since ENs will presumably be small-sized common-of-the-shelf (COTS) infrastructure nodes, as the capacity demand over an EN rises, its CAPEX/OPEX grow exponentially. In fact, EN expenses do not follow the traditional data center cost patterns for this reason [5], [19]. As a result, it will more likely be cheaper to maximize performance and coverage through deploying more ENs, rather than condensing the throughput demand into fewer high-capacity nodes. This reasoning is also supported by the automation levels expected under 5G, as less complex and capacitated ENs could reduce CAPEX/OPEX by being remotely managed and maintained. Nevertheless, given the trade-off regarding ENs number, capacity allocation and throughput requirements, only multi-objective/multi-criteria optimization mechanisms can be used for this particular ecosystem.

C. Traffic Aggregation Points

As explained in Section III-A, any ENPP solution method should be able to handle the demand generated by the traffic aggregation sites mostly in terms of latency and throughput. Consequently, the number of TAPs directly impacts the cost-effectiveness and complexity of the EN placement strategy. Since the number of ENs will scale up to thousands in mid to large-sized cities due to the TAPs amount, solving the ENPP through exact methods is unfeasible. The problem could be simplified by following the MEC approach and co-locating the ENs within existing infrastructure sites, e.g. macro cells, but this is not an efficient solution. A first limitation of this approach is its inability to detect better but still unforeseen locations according to next generation service and demand patterns. Additionally, current EN-capable sites may not answer the specific requirements of 5G use cases and its geo-distribution scheme. Taking this into consideration, the ENPP nature can be assumed to be both discrete (as a potential sites list can be predefined), and continuous (as unforeseen locations must be somehow detected and considered). Overall, the pure continuous approach is mainly intractable and the total of TAPs is expected to be large enough to prevent the use of exact strategies. Hence, pioneering heuristic and meta-heuristic schemes become the best solution candidates.

Fig. 2 presents a possible EC/Cloud multi-tier architecture where more than one EC implementation is deployed according to specific use cases. A mid-tier formed by micro data centers network is in charge of additional edge processing before the remote cloud. In this ecosystem, all nodes in both Tier I and II are considered ENs. Under such architecture as latency, throughput and the number of TAPs may tradeoff with each other, an EN deployment satisfying the three requirements for all 5G scenarios is more likely to be unfeasible due to the involved costs. One of the reasons behind this is that the computation and communication hot spots may not overlap and thus a per-service/per-scenario deployment cannot be used.

Unlike mobile traffic patterns that heavily depend on the end users geo-distribution, the EN placement is tightly coupled to the TAP demand distribution. Consequently, further research is needed in order to characterize and model the TAP traffic behavior for 5G ecosystems. Likewise, it is critical to estimate the EN number and EN type combination for a given demand model, for which recent advances in stochastic geometry theory may be the right solution.

D. Location-dependent costs

When placing an EN, the list of location-dependent costs becomes quite extensive as they go from energy prices and land acquisition to installation expenses. The elements presented in Table II were identified throughout this research as key factors impacting the EN network costs and its placement optimization.

The **power line layout** accounts for the costs of bringing power to the EN site (if needed). Similarly, **network line layout** refers to the cost of bringing networking. These two parameters could effectively push the placement mechanism to: a) reuse existing facilities such as Content Delivery Network Points-of-Presence (CDN-PoPs), Internet Service Providers PoPs (ISP-PoPs) and Central Offices (COs), b) consider the power layout conditions considering the nearest power sources. As a result, significant CAPEX and OPEX reductions could be achieved.

In terms of energy and in close relation to the **power line layout** parameter, a self-sustainable EN location or a green-powered one (ecological energy sources in-use) is preferable. To guarantee this, feasible or available **energy sources** for each EN location should be analyzed. This parameter allows the placement strategy to assess each site regarding its energy supply capabilities. For instance, any location with an in-use ecological power source or capable of using such energy with-

[2] In-place computing/networking resources available for service execution

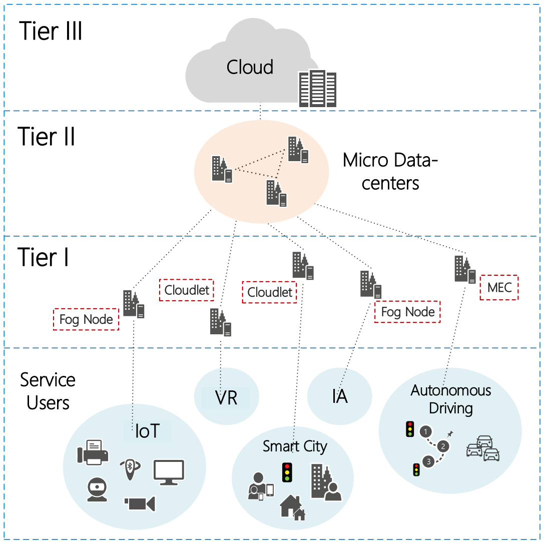

Fig. 2. Cloud/Edge Computing architecture diagram. IoT: Internet of Things, VR: Virtual Reality, IA: Industry Automation

out incurring in high additional expenses, should be ranked higher than other locations with traditional energy supplies.

On the other hand, **build costs** and **land acquisition expenses** are tied to the EN capacity. The former mainly accounts for the cost of installing cooling and power delivery equipment and other support infrastructures, while the latter sums up the costs of renting or buying the required space. Commonly, these expenses are computed in terms of the infrastructure maximum power consumption which is basically determined by the computing resources and the demand. In addition, **site capabilities** evaluates the facility conditions to install computing, networking and storage infrastructure. Consequently, the use of a location with available capacities to deploy an EN (e.g. CDN-PoP) under any business model, such as leasing or hosting, is encouraged. The location-routing nature of the ENPP is taken into account through the **interconnection capabilities**. The communication path between any EN-TG and EN-EN pair is analyzed for each site in order to find those locations where less energy is consumed along the service channel and the lowest capital expenses are needed to ensure interconnection. Any placement method is in charge of ranking all EN potential locations according to already in-place communication infrastructure, IP-capable equipment, radio-wave communications feasibility and interconnection energy consumption.

E. Location restrictions

Not all available sites are suitable for the placement of IT infrastructure. Political, social and environmental conditions should not be overlooked. If a set of potential locations is not identified and the entire geographical area is considered for EN placement, a tessellation or similar method should be applied in order to exclude unfeasible areas and thus optimize the solution search space.

TABLE II
LOCATION-DEPENDENT PARAMETERS

Parameter	Description
Power line layout	Cost of bringing energy to the location
Network line layout	Cost of bringing networking to the location
Energy source	Energy price according to the available power sources
Land acquisition	Cost of renting/buying the required space
Site capabilities	Site conditions to host an EN in terms of computing, storage and networking capacities
Build costs	Cost of deploying the required infrastructure
Interconnection capabilities	EN-TG and EN-EN interconnecting costs

F. Reliability

From a high-level point of view, under the EN placement scope, service availability and reliability depend on budget constraints and site-dependent properties (e.g. natural disaster exposure, network PoPs available, site physical security). Intuitively, disaster-prone areas should be avoided, but a trade-off on this matter should be always kept in mind to avoid overpriced or unfeasible solutions and unmet demand. These elements can be grouped into a parameter conventionally called **site reliability**. Furthermore, in certain scenarios such as mission critical systems, reliability should be ensured through multi-coverage. Therefore, the EN placement mechanisms is forced to deploy additional ENs within the communication range of critical TAPs, according to the latency constraints. Such considerations are thus said to be part of the **coverage reliability** parameter. This factor entails certain particularities as it basically refers to the user demands and not to the site itself. However, if a given user or demand scenario requires coverage from more than one EN, the placement strategy should place in-range additional infrastructure in a different location (in addition to the best location found). Such deployment would imply increasing the overall costs. Therefore, in a first step the placement method should analyze already placed ENs to check whether an existing EN can cover the unsatisfied reliability demands. If such ENs are not found, the placement solution must propose a suitable additional location. A summary of these findings is presented in Table III.

G. Service Area Type

Partitioning and classifying the service areas into urban and rural decreases the execution times of the proposed solving schemes while keeping accuracy, efficiency and performance. Moreover, given the significant difference among **service area type** characteristics, different placement parameters or schemes could be considered accordingly. Rural areas, for instance, are mainly prone to a co-location solution, where ENs are to be deployed in existing communication or computing facilities such as mobile macro cell sites. In contrast, the traffic density and use case mixture in urban and even suburban environments, could force the ENPP solver to check a list of potential sites or the continuous placement space in order to propose EN optimal locations.

TABLE III
RELIABILITY PARAMETERS

Parameter	Description
Site reliability	Site characteristics such as natural disaster exposure and physical security
Coverage reliability	Sensitive use cases where user demands must be satisfied by more than one EN

H. Virtual Network Functions placement

In contrast to the ENPP, the VNF placement problem has been exhaustively tackled [20]. Available VNF placement methods should be carefully considered as they provide valuable hints on how to distribute physical resources. Furthermore, since capacity planning is still an unsolved challenge for a merged 5G/EC architecture, analyzing VNF allocation methods in search for its interrelation with the ENPP resource distribution is mandatory. Certainly, current data center placement guidelines based on energy consumption cannot be followed as a different approach is needed under EC. In such ecosystem, power usage must be optimized based on the whole EN network instead of isolated ENs. Additionally, infrastructure capacity must be planned according to the covered use cases.

IV. PROPOSAL VALIDATION CONSIDERATIONS

Recent work on service infrastructure placement partially supports and validates our findings. In [12], latency, capacity, demand geo-distribution, reliability and a simplified deployment cost model, are used in order to find the best edge locations for computing infrastructure. Authors in [21] propose a latency-constrained method to select the best ENs locations. By comparing two heuristics and an exact method through simulations, it is concluded that location optimization may significantly reduce overall expenses. Additionally, Wang et al. article [17] considers workload efficient distribution and strong latency constraints as core parameters to choose adequate edge server locations. What is more, previous research in [5] has demonstrated the usefulness of a thorough site selection planning for service infrastructure deployment and its direct impact on data center deployment and operation costs.

The references cited throughout this paper provide a valid groundwork for a placement parameter set. However, a lack of criteria completeness could be observed when attempting to effectively place ENs under 5G. For instance, to the best of our knowledge, Section III-A is the first step into defining which delay values should be be considered when placing an EN. Reliability has been mainly overlooked in most placement research, although the foreseeable EN deployment density and 5G use case scenarios pose complex requirements in this topic. Additionally, service area type is a novel parameter proposal that directly affects CAPEX/OPEX, along with analyzing VNFs placing methods.

V. CONCLUSION

The EC deployment for next generation 5G networks requires innovative schemes and solutions. This paper sets a starting point for the EN placement optimization towards a feasible 5G-EC ecosystem. By defining a potential list of parameters to solve the ENPP, the groundwork for a cost-effective solution strategy has progressed further. Future research should focus on a deep understanding of capacity planning requirements and guidelines to ensure the baseline for a joint solution to the EN capacity and site selection problem.

REFERENCES

[1] S. Zhang, X. Xu, Y. Wu, and L. Lu, "5G: Towards energy-efficient, low-latency and high-reliable communications networks," in *2014 IEEE International Conf. on Comm. Systems*, 2014, pp. 197–201.

[2] I. Parvez *et al.*, "A survey on low latency towards 5g: RAN, core network and caching solutions," *arXiv:1708.02562 [cs]*, 2017.

[3] GS MEC-IEG ETSI, "Mobile edge computing; market acceleration; MEC metrics best practice and guidelines," ETSI, Tech. Rep. 006 V1.1.1, 2017.

[4] H. Tanaka *et al.*, "Multi-access edge computing: A survey," *Journal of Information Processing*, vol. 26, no. 0, pp. 87–97, 2018.

[5] I. Goiri *et al.*, "Intelligent placement of datacenters for internet services," in *2011 31st ICDCS*. IEEE, 2011, pp. 131–142.

[6] W. Zhang *et al.*, "Infrastructure deployment and optimization of fog network based on MicroDC and LRPON integration," *Peer-to-Peer Networking and Applications*, vol. 10, no. 3, pp. 579–591, 2017.

[7] A. B. Arabani and R. Z. Farahani, "Facility location dynamics: An overview of classifications and applications," *Computers & Industrial Engineering*, vol. 62, no. 1, pp. 408–420, 2012.

[8] S. Wang and C. Ran, "Rethinking cellular network planning and optimization," *IEEE Wireless Comm.*, vol. 23, no. 2, pp. 118–125, 2016.

[9] S. Barbarossa, S. Sardellitti, and P. Di Lorenzo, "Communicating while computing: Distributed mobile cloud computing over 5G heterogeneous networks," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 45–55, 2014.

[10] M. Yannuzzi *et al.*, "A new era for cities with fog computing," *IEEE Internet Computing*, vol. 21, no. 2, pp. 54–67, 2017.

[11] I. Gravalos *et al.*, "Efficient gateways placement for internet of things with QoS constraints," in *2016 IEEE GLOBECOM*, 2016, pp. 1–6.

[12] H. Yin *et al.*, "Edge provisioning with flexible server placement," *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, no. 4, pp. 1031–1045, 2017.

[13] B. Yang *et al.*, "Cost-efficient NFV-enabled mobile edge-cloud for low latency mobile applications," *IEEE Transactions on Network and Service Management*, vol. 15, no. 1, pp. 475–488, 2018.

[14] A. Mukherjee, "Energy efficiency and delay in 5G ultra-reliable low-latency communications system architectures," *IEEE Network*, vol. 32, no. 2, pp. 55–61, 2018.

[15] V. G. Nguyen *et al.*, "SDN/NFV-based mobile packet core network architectures: A survey," *IEEE Comm. Surveys Tutorials*, vol. 19, no. 3, pp. 1567–1602, 2017.

[16] J. Zhang *et al.*, "Mobile edge computing and field trial results for 5G low latency scenario," *China Communications*, vol. 13, pp. 174–182, 2016.

[17] S. Wang *et al.*, "Edge server placement in mobile edge computing," *Journal of Parallel and Distributed Computing*, 2018.

[18] M. Emara, M. C. Filippou, and D. Sabella, "MEC-aware cell association for 5G heterogeneous networks," in *2018 IEEE Wireless Comm. and Networking Conference Workshops (WCNCW)*, 2018, pp. 350–355.

[19] L. A. Barroso, J. Clidaras, and U. Holzle, *The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines*, 2nd ed., ser. Synthesis lectures on computer architecture. Morgan & Claypool, 2013, no. 6.

[20] X. Li and C. Qian, "A survey of network function placement," in *2016 13th IEEE CCNC*. IEEE, 2016, pp. 948–953.

[21] A. Santoyo-González and C. Cervelló-Pastor, "Latency-aware cost optimization of the service infrastructure placement in 5G networks," *Journal of Network and Computer Applications*, vol. 114, pp. 29–37, 2018.