CHALMERS
UNIVERSITY OF TECHNOLOGY

End-to-End Provisioning of Latency and Availability Constrained 5G Services

Citation for the original published paper (version of record):
Lashgari, M., Wosinska, L., Monti, P. (2021). End-to-End Provisioning of Latency and Availability Constrained 5G Services. IEEE Communications Letters, 25(6): 1857-1861.
http://dx.doi.org/10.1109/LCOMM.2021.3063262

End-to-End Provisioning of Latency and Availability Constrained 5G Services

Maryam Lashgari, *Member, IEEE,* Lena Wosinska, *Senior Member, IEEE,* and Paolo Monti, *Senior Member, IEEE*

***Abstract*—We address a key challenge of 5G networks by proposing a strategy for the resource-efficient and end-to-end allocation of compute and connectivity resources in a dynamic 5G service provisioning scenario, such that the service latency and availability requirements are guaranteed. Our heuristic algorithm shows that resource efficiency is significantly improved by processing services in the large core data centers (DCs) with a rich amount of compute resources and exploiting the benefits of traffic grooming over the metro and core fiber links. Moreover, our resource-efficient provisioning algorithm avoids possible violation of the service availability requirements caused by reaching the central DC locations by adding backup connectivity resources. Our simulation results demonstrate a resource efficiency improvement reflected by lowering the service blocking probability by up to four orders of magnitude compared to the conventional service provisioning methods utilizing distributed small DCs.**

***Index Terms*—Optical networking, 5G, cloud computing, network control and management, service provisioning, latency, availability, backup connectivity.**

I. Introduction

5G networks need to support services with stringent latency, availability, and capacity requirements by connecting end-users to the data center (DC) locations where the application servers (ASs) (responsible for executing the service-specific applications) are deployed [1]. In this regard, there is a need for service provisioning strategies that assure meeting service requirements while using radio, connectivity, and compute resources efficiently.

The location of the AS is crucial. The ASs that are deployed close to the users, i.e., in edge/access DCs offer better latency and availability performance compared to the ASs located in more centralized DCs (i.e., in metro and/or core). However, edge DCs have (on average) limited compute resources and host fewer ASs, while metro/core DCs have more compute resources available. The metro/core DCs are also reachable through higher tier transport network (TN) segments with higher link capacity, which allows for efficient multiplexing and utilization of network connectivity resources. As a result, processing services in larger and more centralized DC locations have the potential to improve the efficiency in which connectivity and compute resources are utilized in a 5G network infrastructure.

However, to take advantage of the benefits of centralized processing, one needs to make sure that the service latency and availability requirements are guaranteed. From a latency perspective, centralized service processing is allowed only up to a certain level, i.e., long paths towards DCs in higher network tiers will eventually violate the latency constraint. On the other hand, the low availability performance associated with such long paths will be improved by adding redundant connectivity resources where needed. As a result, as long as the latency constraints are met, adding backup connectivity resources (if/where needed) makes it possible to reach more central DCs without violating the service availability constraint. This intuition was effective in reducing the deployment cost of 5G communication infrastructures, as shown in [2], where the authors proposed a network design strategy to place ASs at DC locations as central as possible by adding backup connectivity resources needed to guarantee the required service availability level. Their results showed up to 74% savings of the infrastructure deployment costs. On the other hand, using the same intuition (i.e., centralized processing aided by backup connectivity) while operating a 5G network infrastructure might lead to an increased amount of connectivity resources that (on average) are needed to provision a given service. This, in turn, might limit the number of services a provider can accommodate in its infrastructure. For this reason, it is still an open question of whether or not centralized service processing is beneficial for dynamic provisioning of services with strict latency and availability requirements.

There are several published works addressing the service provisioning problems with latency- and/or availability-guarantee in 5G networks. The authors in [3] proposed a restoration-based survivable strategy leveraging the cloud service relocation concept. Their target was to minimize the average service downtime and the number of relocated cloud services. The work in [4] proposed a provisioning solution for services with specific latency requirements. The objective was minimizing the usage of fiber, processing, and storage resources while deploying service applications in hierarchically distributed DCs. The authors in [5] considered a wireless-optical converged network and proposed a slice provisioning algorithm whose aim is to maximize the efficiency in which both optical and wireless resources were used while meeting the delay and bandwidth requirements of the services. Despite the many important challenges addressed in the literature so far, currently, there are no works that focused on optimizing resource efficiency while provisioning 5G services with both latency and availability guarantee.

This work proposes an end-to-end provisioning strategy for 5G services, referred to as resource-efficient provisioning (REP), that addresses the service latency and availability constraints concurrently. The method presented in the paper aims at maximizing the resource efficiency of both the connectivity and compute resources available in the 5G network infrastructure. REP leverages the intuition presented in [2] where services should be processed at DC locations as central as possible (while satisfying the latency constraints) to take advantage of *(i)* the multiplexing capabilities of the connectivity

M. Lashgari, L. Wosinska, and P. Monti are with the Electrical Engineering Department, Chalmers University of Technology, Gothenburg, Sweden, e-mail: ({maryaml, wosinska, mpaolo}@chalmers.se).

links in the metro and core network segments, and *(ii)* the abundant compute resources available in the metro/core DC locations compared to the access DCs. This is made possible by adding backup connectivity resources whenever the path connecting the end-user to the DC location, where the service-specific AS is running, does not meet the service availability requirements.

In the proposed provisioning algorithm, resource efficiency is maximized by selecting a DC location, a connectivity path, and (optionally) a backup path for each service request such that a metric measuring how both the connectivity and compute resources are utilized is minimized. The REP strategy is tested using two use cases, one considering a 5G service with strict latency and availability constraints, and another one with a 5G service with more relaxed latency and availability requirements. Simulation results show that, compared to the conventional provisioning strategies (i.e., without the ability to provide a backup path where needed, leading to the deployment of services only at the edge of the network because of the service availability constraints), REP improves the service blocking performance by up to four orders of magnitude in the presence of 5G service with strict latency and availability requirements and by up to two orders of magnitude when the service latency and availability constraints are relaxed.

II. System Architecture, Latency and Availability Models

This work assumes the network architecture presented in Fig. 1. The TN is a packet over an optical network, where the optical infrastructure uses a wavelength division multiplexing (WDM) technology. This latter assumption is made for simplicity, and the service provisioning strategy proposed in the paper can be easily adapted to elastic optical networks. This work also assumes a software-defined networking-based control architecture where an orchestration layer has a global view of the infrastructure while managing connectivity (wireless and TN) and compute resources via their respective controllers [6].

A service request consists of one or more user equipments (UEs) (connected to one of the wireless access points (APs)) requiring the deployment of a service-specific AS in one of the available DC locations. Upon a service request, two types of resources need to be provisioned: *(i)* compute (i.e., deployment of the AS) and *(ii)* connectivity (i.e., for the path connecting the DC, where the AS is running, and the AP to which the UEs are connected to).

UEs are connected to APs through wireless links, while APs are dual-homed to their respective access edge (AE) nodes through dedicated fiber links. The traffic that goes over the TN is of the midhaul type (i.e., functional split option 2 is selected, whereas the distributed and centralized units are deployed at APs and the DC locations, respectively [7]). The TN consists of three tiers (i.e., access, metro, and core) working at different transmission rates. The access and metro segments are ring-based, while the core network has a mesh topology. This is in line with the architecture presented in [8] which shows good performance in terms of cost-efficiency. The access rings are connected to their respective metro rings via metro nodes (MNs), while the metro rings are connected to the core mesh via metro-core edge (ME) nodes (Fig. 1). Traffic grooming (i.e., to multiplex low-rate flows into higher rate ones) is performed only at the ME and MN locations.

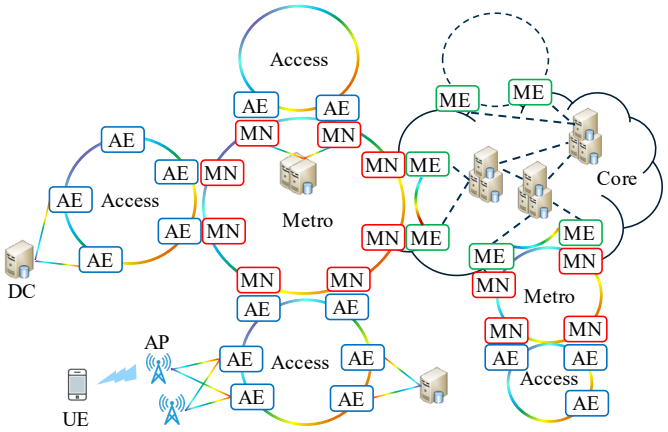

Fig. 1. Network architecture with three tiers (i.e., access, metro, and core). UEs connect to the infrastructure via wireless APs. Services can be processed at DCs placed in any of the network segments. AE: access edge, MN: metro node, ME: metro-core edge, UE: user equipment.

Three types of DCs are available in the network, i.e., core, metro, and access. The DC locations in each access ring are dual-homed to the AE nodes, while the DCs in each metro ring are dual-homed to the MNs. The DCs in the core network are dual-homed to the ME nodes. The core DCs are assumed to have more compute capabilities compared to the metro ones, which, in turn, are equipped with more compute resources than the access DCs.

This work considers a dynamic provisioning scenario in which services are requested at different points in time. A service request j comes with specific requirements in terms of end-to-end (E2E) latency ($l_{(j,e2e)}$) (i.e., the latency experienced by the service data flow), E2E availability ($a_{(j,e2e)}$), data rate (r_j), and compute resources (s_j). The values of $l_{(j,e2e)}$ and $a_{(j,e2e)}$ are derived from [9] and [2] and are computed as follows:

$$l_{(j,e2e)} = l_{\text{UE}} + l_{\text{RAN}} + l_{\text{sw}} + l_{\text{AS}} + l_{\text{prop}}, \tag{1}$$

$$a_{(j,e2e)} = a_{\text{UE}} \times a_{\text{RAN}} \times a_{\text{TN}} \times a_{\text{AS}}, \tag{2}$$

where l_{UE}, l_{RAN}, l_{sw}, l_{AS}, and l_{prop} represent the latency values of the UE, of the radio access network (RAN), of a switching node in the TN (i.e., due to grooming operations), of the AS (i.e., due to processing), and the propagation latency in the TN, respectively. l_{prop} is defined as the ratio between the distance traversed by the path connecting an AP and a DC location (i.e., σ) and the speed of light propagating into the fiber (i.e., v). For the availability computation, a_{UE}, a_{RAN}, a_{TN}, and a_{AS} represent the availability values of the UE, RAN, TN, and AS, respectively. The availability value of a single path in the TN (i.e., without any backup resources) is calculated as:

$$a_{\text{TN}} = \prod_{n=1}^{N} (1-\mu_n) \times \prod_{e=1}^{N+1} (1-\rho_e \times \zeta_e), \tag{3}$$

where μ_n and ρ_e are the values of the failure probability of node n and link e (i.e., per [km]), respectively. ζ_e is the length of link e, and N is the number of nodes in the path from the AP to the DC where the AS of a given service is deployed.

The main idea behind this work is to maximize the efficiency in which the infrastructure resources (i.e., both connectivity and compute) are used. This is achieved by trying to deploy an AS in the most central DC location as possible

to take advantage of *(i)* a large amount of compute resources that high tier DC locations have, and *(ii)* to fully exploit the benefits of traffic grooming over the metro and core fiber links. Centralizing the deployment of AS has to be done while satisfying the service latency and availability requirements. The latency is a function of the distance between the AP and the DC location, and of the processing time (i.e., at the user side, at switching nodes along the path in the TN and in the AS). The availability can be improved by adding (if needed) backup connectivity resources to the path between the AP and the candidate DC location where the AS is deployed. The next section presents a service provisioning strategy that leverages this intuition.

III. Resource-Efficient Service Provisioning Strategy

This section describes the resource-efficient provisioning (REP) strategy proposed in the paper. The problem solved by REP is formulated as follows. Given a service request j originating at an AP, i.e. δ, with E2E latency, E2E availability, compute, and data rate requirements defined by $\left(l_{(j,e2e)}, a_{(j,e2e)}, s_j, r_j\right)$, REP selects a DC location d (i.e., where the AS of j is deployed), a connectivity path p, and (optionally) a backup path b_p between δ and d such that the cost parameter $c_{(j,d,p,b_p)}$ is minimized. We derived the resource consumption metric $c_{(j,d,p,b_p)}$ as:

$$c_{(j,d,p,b_p)} = \beta \cdot \frac{s_j}{m_d} + \alpha \cdot \left(\sum_{e \in \mathcal{E}_{j,d,p}} \frac{\eta_{j,e}}{w_e} + \sum_{e \in \mathcal{E}_{j,d,b_p}} \frac{\eta_{j,e}}{w_e} \right) \quad (4)$$

where m_d is the overall compute capacity available at d, $\eta_{j,e}$ is the number of wavelengths required by j over link e, w_e is the overall wavelength capacity of link e, while $\mathcal{E}_{j,d,p}$ and $\mathcal{E}_{j,d,b_p}$ are the set of links comprising p and b_p, respectively. If p meets the availability requirement of j without the need for a backup path, then $\mathcal{E}_{j,d,b_p} = \emptyset$. The resource consumption metric described in (4) is designed to reflect the intuition that since compute resources in the access are scarce compared to the ones in the metro and core network segments, they should be used wisely, to avoid creating unnecessary bottlenecks. The first term in (4) is used to encourage using compute resources in DCs equipped with more resources, while the second and third terms are used to choose connectivity paths over the links that have more wavelength resources available. α and β are two weighting factors that can be tuned according to the amount of the connectivity and compute resources available in the network infrastructure.

Upon the arrival of a new service request j, the algorithm works as follows (Algorithm 1). For each candidate $d \in \mathcal{D}$ (i.e., the set of DCs whose free compute resources are equal or greater than s_j), REP looks at all candidate paths (i.e., g) between d and δ. These paths are pre-computed using k-shortest path algorithm and are stored in set $\mathcal{G}_{\delta,d}$. If g meets the $l_{(j,e2e)}$ and $a_{(j,e2e)}$ requirements of j, and if g has enough free connectivity resources (i.e., $\geq \eta_{j,e}$) the provisioning solution (d, g) is added to the set of possible provisioning solution options (Q) (step 7). If, on the other hand, $l_{(j,e2e)}$ is not met, REP checks the next candidate path in $\mathcal{G}_{\delta,d}$. If $l_{(j,e2e)}$ is met but $a_{(j,e2e)}$ is not, g is added to the set of paths for which the option of adding a backup can be considered, i.e., ($\mathcal{P}_{\delta,d}$), under the assumption that $l_{(j,e2e)}$ and $a_{(j,e2e)}$ will not be violated (step 10). In step 14, REP checks, one by one, all paths p stored in $\mathcal{P}_{\delta,d}$ to see if a suitable backup path b_p between δ and d can be found. For each $p \in \mathcal{P}_{\delta,d}$, the algorithm examines a number of node disjoint options (i.e., pre-computed by using the k-shortest path algorithm) stored in set $\mathcal{M}_{\delta,d,p}$. If b_p meets $l_{(j,e2e)}$ and $(p+b_p)$ meets $a_{(j,e2e)}$, then b_p is added to the set of candidate backup paths for p (i.e., $\mathcal{B}_{\delta,d,p}$) (step 17). In step 20, REP checks if p and $b_p \in \mathcal{B}_{\delta,d,p}$ have enough free connectivity resources. If yes, (d, p, b_p) is added to the set of possible provisioning solution options (Q). Once all the DC options that can accommodate j have been checked, the element in (Q) which minimizes $c_{(j,d,p,b_p)}$ in (4) is chosen as the final solution (step 26). If set (Q) is empty, service j is rejected (step 28).

Algorithm 1 resource-efficient provisioning (REP) strategy

1: Given a service j at AP, δ with $\left(l_{(j,e2e)}, a_{(j,e2e)}, s_j, r_j\right)$ requirements
2: **for** $d \in \mathcal{D}$ **do**
3: **for** $g \in \mathcal{G}_{\delta,d}$ **do**
4: **if** $l_{(j,e2e)}$ is met
5: **if** $a_{(j,e2e)}$ is met
6: **if** g has free capacity
7: Add $q = (d, g)$ to Q and **break**
8: **end if**
9: **else**
10: Add g to $\mathcal{P}_{\delta,d}$
11: **end if**
12: **end if**
13: **end for**
14: **for** $p \in \mathcal{P}_{\delta,d}$ **do**
15: **for** $b_p \in \mathcal{M}_{\delta,d,p}$ **do**
16: **if** b_p meets $l_{(j,e2e)}$ and $(p+b_p)$ meets $a_{(j,e2e)}$
17: Add b_p to $\mathcal{B}_{\delta,d,p}$
18: **end if**
19: **end for**
20: **if** p and $b_p \mid b_p \in \mathcal{B}_{\delta,d,p}$ have free resources
21: Add (d, p, b_p) to Q and **break**
22: **end if**
23: **end for**
24: **end for**
25: **if** $Q \neq \emptyset$
26: Return $q \in Q \mid c_{(j,d,p,b_p)}$ is minimized
27: **else**
28: Return "null"
29: **end if**

The worst case computational complexity of the REP algorithm is computed as $O\left(|\mathcal{D}| \times |\mathcal{G}_{\delta,d}| \times |\mathcal{B}_{\delta,d,p}| + |Q|\right)$.

IV. Simulation Results

This section presents the performance evaluation results of the REP strategy. The results are obtained via an ad-hoc, Python-based, event-driven simulator which implements the system model described in Sec. II, and the provisioning strategy presented in Sec. III.

In the simulations, we assume to have 2 metro rings and 3 access rings connected to each metro ring. Each access ring consists of 10 AE nodes, while each metro ring comprises 8 MNs. The number of DCs connected to each access and metro ring is equal to 30% and 20% of the total number of

TABLE I
USE CASES UNDER EXAM: AR AND V2X WITH THEIR REQUIREMENTS AND NUMBER OF ASSUMED USERS [9], [11]–[14].

	V2X	AR
Latency [ms]	10	100
Availability [%]	99.99	99.9
Connectivity [Mbps]	25	10
Compute [CU]	0.2	0.02592
Number of users	(3, 30)	(10, 50)

AEs and MNs (rounded up to the next integer), respectively. The AE nodes or MNs to which each DC is dual-homed are chosen randomly with a uniform distribution. The core network consists of 4 ME nodes (i.e., 2 ME nodes connected to each metro ring) forming a square topology. There is 1 core DC connected to all of the ME nodes. The number of APs connected to each access ring is equal to 30% of the number of AEs in the access ring. The AEs to which an AP is dual-homed are chosen randomly with a uniform distribution. The link length, i.e., the value of ζ_e, in the core, metro, and access rings is 500, 70, and 3 [km], respectively.

The value of m_d, expressed in computing units [CUs], for each DC in the access, metro, and core segment is 5, 25, and 125 [CUs], respectively. The number of wavelengths, i.e., the value of w_e, in the core, and metro links is assumed to be 80, while it is 10 in access links. A wavelength in the core, metro, and access segment operates at a transmission rate of 100, 10, and 1 [Gbps], respectively. Traffic grooming is performed only at ME and MN locations. Each time grooming is performed a switching latency (i.e., l_{sw}) is introduced and added to the total latency budget of the service. We assume that v=2 × 10^8 [m/s], l_{RAN}=3 [ms], l_{sw}=0.2 [ms] [9], and l_{UE}=l_{AS}=0 [ms], while ρ_e=10^{-5}/[km] ($\forall e$) [10], and μ_n=10^{-6} ($\forall n$) [9], and a_{UE}=a_{RAN}=a_{AS}=1.

The service request arrivals follow a Poisson process. The inter-arrival time is exponentially distributed with a rate of λ. The service holding time is exponentially distributed with a mean value equal to 24 time units [TUs].

Two use cases are considered in the study. In each one of them, a specific 5G service is provisioned, i.e., vehicle-to-X for short term environment modelling (V2X-STEM) (e.g., sensor sharing) and augmented reality (AR) for medical applications (Table I). They are two examples of 5G services with strict (i.e., V2X) and relaxed (i.e., AR) latency and availability requirements. On the other hand, the REP algorithm can be applied in the presence of any type of service. The number of the users considered for each service request is chosen uniformly within the ranges shown in Table I, leading to different requirements for connectivity and compute resources for each service instance. The values of weighting factors α and β defined in (4) are tuned according to the operator's needs. In this study we assume $\alpha = 0.1$ and $\beta = 1$.

The performance of REP is evaluated against a benchmark strategy referred to as no path protection (NPP). NPP is derived from [4] where no backup resources can be added along the path connecting the AP and the DC location where the service-specific AS might be deployed. Two performance metrics are used, service blocking probability (BP) and average compute utilization (AVCU). The former is measured as the ratio between the number of rejected service requests and the total number of processed service requests. The latter measures the amount of the total available compute resources

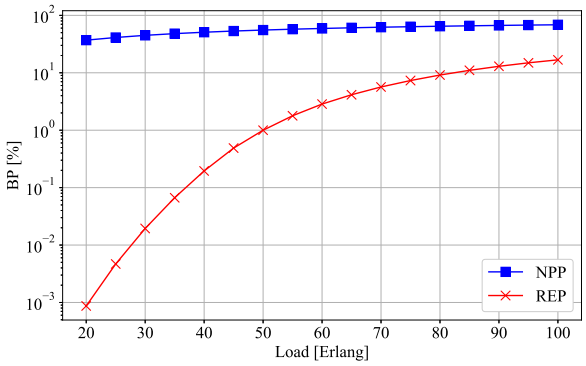

Fig. 2. V2X use case: service BP vs. Load ($a_{j,e2e}$ = 99.99%).

that is used (on average) in a given network segment (i.e., access, metro, or core). More formally, AVCU is defined as:

$$AVCU = \frac{1}{|\mathcal{I}|} \sum_{d \in \mathcal{I}} \sum_{j \in \mathcal{S}_d} \frac{h_j \times s_j}{T \times m_d} \tag{5}$$

where $\mathcal{I}$ is the set of DCs in a given network segment (i.e., access, metro, or core), $\mathcal{S}_d$ is the set of all services deployed in DC d during the simulation time T, and h_j is the holding time of service j. The results presented in this section are the average over a large number of experiments (i.e., each one with T = 296000 [TU]) enough to obtain a confidence interval for the value of the service BP less or equal than 2.8% with a 95% of confidence level.

Figure 2 compares the service BP values of the V2X service use case as a function of the network load for both the NPP and the REP approaches. Overall, REP outperforms NPP under all load conditions (four orders of magnitude gain compared to NPP for low values of load). Since V2X has stringent availability requirements, the NPP strategy deploys ASs only at access DC locations which, in turn, have limited compute resources. This is because reaching metro and core DCs would violate the service availability constraint. On the other hand, REP can add backup connectivity resources when needed and, as a result, can deploy ASs in metro and core DCs where compute resources are more abundant. This intuition is confirmed by the results presented in Fig. 3 where the values of AVCU for the V2X use case are compared as a function of the network load. To deploy an AS, NPP relies only on access compute resources. REP, on the other hand, in addition to the access DC resources, can also consider metro and core DC locations. Overall, the value of AVCU increases with increasing load values. This is in line with the intuition that with more services in operation, more compute resources will be used in each DC.

Figures 4 and 5 present similar results to the ones in Figs. 2 and 3 but for the AR use case, which has less stringent availability and latency requirements compared to the V2X use case. REP strategy gains more than two orders of magnitude in terms of service BP compared to NPP (Fig. 4) thanks to the possibility of deploying ASs at more central DC locations. The reduced gain compared to the V2X use case is because NPP can now deploy AS in both access and metro DCs, thanks to the lower availability requirements associated with the AR service. Most of the service BP gain of REP comes from its

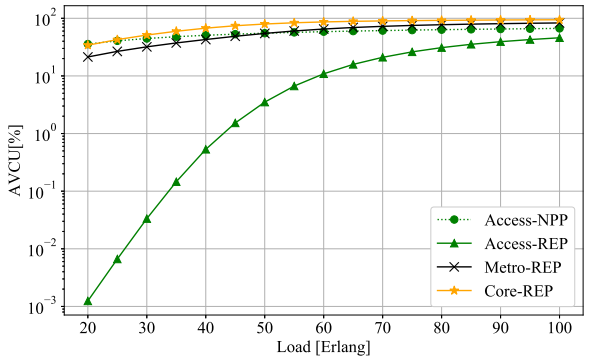

Fig. 3. V2X use case: AVCU vs. Load with ($a_{j,e2e}$ = 99.99%).

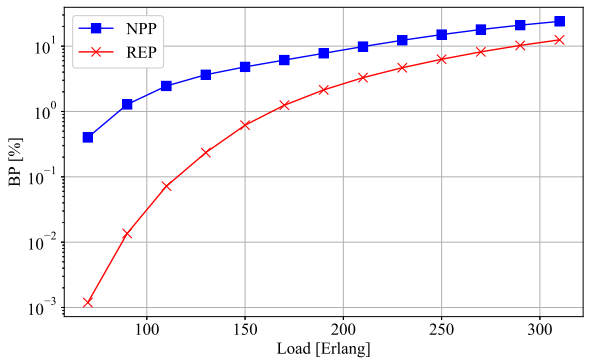

Fig. 4. AR use case: service BP vs. Load ($a_{j,e2e}$ = 99.9%).

ability to use compute resources deployed at the core DCs (Fig. 5).

V. Conclusions

This work presented a resource-efficient, dynamic service provisioning strategy that allocates compute and connectivity resources in a 5G infrastructure such that the end-to-end service latency and availability requirements are met. The proposed heuristic algorithm, referred to as REP, leverages the possibility of adding backup connectivity resources along the path from the AP (i.e., where an end-user is connected to) to the DC location, where the service-specific AS is deployed. This allows services to be processed at the most central DC locations as possible (i.e., meeting the service latency constraints), an option that allows operators to exploit *(i)* the multiplexing gains offered by metro and core network segments, and *(ii)* the abundant compute resources available in metro and core DCs. This, in turn, improves the resource efficiency of the deployed infrastructure. The performance of the REP algorithm, evaluated by simulating two representative 5G use cases, shows how the centralized service processing leads to up to four orders of magnitude improvements in terms of service blocking probability compared to conventional dynamic service provisioning strategies.

Acknowledgment

This work was funded by the MSCA-ITN project 5G STEP FWD with funding from the European Union's Horizon 2020 research under grant agreement number 722429. The authors would also like to thank C. Natalino and L. M. Contreras-Murillo for useful discussions.

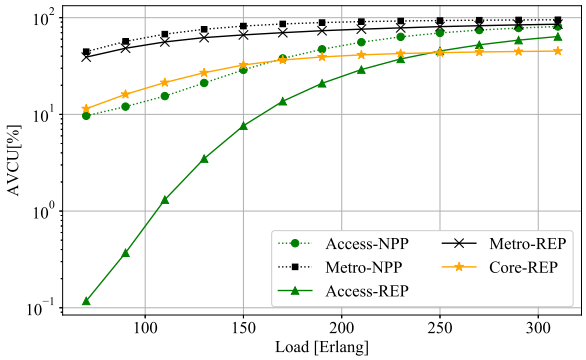

Fig. 5. AR use case: AVCU vs. Load ($a_{j,e2e}$ = 99.9%).

References

[1] T. Doukoglou, V. Gezerlis, K. Trichias, N. Kostopoulos, N. Vrakas, M. Bougioukos, and R. Legouable, "Vertical industries requirements analysis targeted KPIs for advanced 5G trials," in *European Conference on Networks and Communications (EuCNC)*, 2019, pp. 95–100.

[2] M. Lashgari, C. Natalino, L. M. Contreras, L. Wosinska, and P. Monti, "Cost benefits of centralizing service processing in 5G network infrastructures," in *Asia Communications and Photonics Conference (ACPC)*, 2019, p. M3C.2.

[3] C. N. da Silva, L. Wosinska, S. Spadaro, J. C. W. A. Costa, C. R. L. Frances, and P. Monti, "Restoration in optical cloud networks with relocation and services differentiation," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 8, no. 2, pp. 100–111, 2016.

[4] R. F. Vieira, P. H. Alves Pereira, and D. L. Cardoso, "Resource allocation optimization for hierarchical cloud data centers," in *4th International Conference on Cloud Computing Technologies and Applications (Cloudtech)*, 2018, pp. 1–6.

[5] M. Mosahebfard, J. Vardakas, K. Ramantas, and C. Verikoukis, "SDN/NFV-based network resource management for converged optical-wireless network architectures," in *21st International Conference on Transparent Optical Networks (ICTON)*, 2019, pp. 1–4.

[6] M. R. Raza, M. Fiorani, A. Rostami, P. Öhlen, L. Wosinska, and P. Monti, "Dynamic slicing approach for multi-tenant 5G transport networks [invited]," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 10, no. 1, pp. A77–A90, 2018.

[7] L. M. P. Larsen, A. Checko, and H. L. Christiansen, "A survey of the functional splits proposed for 5G mobile crosshaul networks," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 146–172, 2019.

[8] B. Skubic and I. Pappa, "Energy consumption analysis of converged networks: Node consolidation vs metro simplification," in *2013 Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference (OFC/NFOEC)*, 2013, pp. 1–3.

[9] NGMN Alliance, "5G extreme requirements: End-to-end considerations," White paper, 2018, version 2.0.

[10] O. Dobrijevic, C. Natalino, M. Furdek, H. Hodzic, M. Dzanko, and L. Wosinska, "Another price to pay: An availability analysis for SDN virtualization with network hypervisors," in *10th International Workshop on Resilient Networks Design and Modeling (RNDM)*, 2018, pp. 1–7.

[11] M. Boban, A. Kousaridas, K. Manolakis, J. Eichinger, and W. Xu, "Use cases, requirements, and design considerations for 5G V2X," 2017. [Online]. Available: http://arxiv.org/abs/1712.01754

[12] NGMN Alliance, "Perspectives on vertical industries and implications for 5G," White paper, 2016, version 2.0.

[13] A. Moubayed, A. Shami, P. Heidari, A. Larabi, and R. Brunner, "Edge-enabled V2X service placement for intelligent transportation systems," *IEEE Transactions on Mobile Computing*, pp. 1–1, 2020.

[14] A. Hmaity, M. Savi, L. Askari, F. Musumeci, M. Tornatore, and A. Pattavina, "Latency- and capacity-aware placement of chained virtual network functions in FMC metro networks," *Optical Switching and Networking*, vol. 35, p. 100536, 2020.