Cost Benefits of Centralizing Service Processing in 5G Network Infrastructures

Downloaded from: https://research.chalmers.se, 2026-01-23 23:59 UTC

Citation for the original published paper (version of record):

Lashgari, M., Natalino Da Silva, C., M. Contreras, L. et al (2019). Cost Benefits of Centralizing Service Processing in 5G Network Infrastructures. Optics InfoBase Conference Papers, Part F138-ACPC 2019(F138-ACPC 2019)

N.B. When citing this work, cite the original published paper.

research.chalmers.se offers the possibility of retrieving research publications produced at Chalmers University of Technology. It covers all kind of research output: articles, dissertations, conference papers, reports etc. since 2004. research.chalmers.se is administrated and maintained by Chalmers Library

(article starts on next page)

Cost Benefits of Centralizing Service Processing in 5G Network Infrastructures

M. Lashgari[1], C. Natalino[1], L. M. Contreras[2], L. Wosinska[1], P. Monti[1]

[(1)]*Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, Sweden.*
[(2)]*Telefónica I+D, Madrid, Spain.*
maryaml@chalmers.se

Abstract: We assess the benefits of centralizing service processing in a few high-scale data center locations within an operator infrastructure. Results show up to 74% less cost while provisioning latency and availability constrained services. © 2019 The Author(s)

OCIS codes: (060.4510) Optical communications, (060.4250) Networks.

1. Introduction

In 5G networks, services will require a combination of radio, transport, and storage/compute resources and will have service level agreement (SLA) constraints defined in terms of capacity, latency, and availability. Many of these services will rely heavily on virtualized network functions that can be placed at data centers (DCs) located throughout the communication infrastructure. Operators can leverage upon this degree of freedom by ensuring, from one side, that the SLA requirements of each service are always met, and, from the other side, that their communication infrastructure is used cost-effectively.

From a cost perspective, there are two main reasons why it might be beneficial for an operator to deploy services at large and high-scale DCs. First, thanks to the economy of scale, the cost of processing services in a large DC is lower compared to the cost of deploying a service in a small DC [1]. Second, large DCs are usually reachable via high tier transport network (TN) segments (e.g., regional or national) with high bit rate channels that allow the multiplexing of a large number of services over few connectivity resources [2]. However, from a quality of service perspective, placing services in high-scale DCs increases latency (i.e., services are deployed over long connectivity paths) and degrades reliability performance (i.e., services traverse more components, which increases the risk of being affected by failures). While nothing can be done to improve latency besides choosing a DC close to the end-user, the availability of a service deployed over a long connectivity path can be improved by adding protection (i.e., redundant) resources. Therefore, as long as the latency constraints are met, protection techniques can be used to allow services to be deployed at high-scale DCs. In turn, this will allow operators to fully exploit both the cost efficiency of high-scale DCs and the multiplexing gains of high tiers TN segments. Nonetheless, the use of redundant TN resources increases the infrastructure cost. For this reason, the impact of this extra expenditure needs to be quantified.

This paper investigates the trade-offs between the cost benefits of centralizing services at few high-scale DCs and the cost required for protecting the TN resources to meet the availability requirement of some of the deployed services. Four use cases (UCs) are considered, each one with different latency and availability constraints. Numerical results show that when the latency constraints allow for services to be placed at high-scale DCs, the use of redundant TN resources to improve service availability yields up to 74% savings in terms of the overall communication infrastructure cost.

2. Latency, Availability and Infrastructure Cost Computation

This work considers the architecture described in [3] and illustrated in Fig. 1. In this architecture, user equipments (UEs) connect to access points (APs) through wireless links. APs are connected to DCs through an optical TN. DCs host the application servers (ASs) supporting the execution of applications. DCs are placed in the network at selected locations such that the service latency and availability constraints can be met. With DCs close to APs, shorter AP-AS paths can be established over the TN, but each AS supports a low number of APs. If DCs are placed far from the APs, longer TN paths are required, but each AS can handle a large number of APs. We assume a TN composed of three segments. The *local* segment connects the AP to the local aggregation point through dedicated

Fig. 1: Network architecture.

fiber connections. At the local aggregation point, the traffic is aggregated and sent to the *province* segment. In turn, at the province aggregation point, traffic is aggregated and sent to the *regional* segment and so on until a DC location is reached. Inside each TN segment, the traffic is switched all optically, i.e., packet switching is performed only at the nodes at the edge of each TN segment.

$$l_{tot} = l_{UE} + l_{RAN} + l_{sw} + l_{prop} + l_{AS} \quad (1) \qquad a_{e2e} = a_{UE} \times a_{RAN} \times a_{TN} \times a_{AS} \quad (2)$$

$$a_{TN} = \prod_{i=0}^{n}(1 - f_i^{node}) \times \prod_{j=0}^{n+1}(1 - f_j^{link} \times d_j) \quad (3)$$

The end-to-end service latency (l_{tot}) is expressed by Eq. (1), where l_{UE}, l_{RAN}, l_{sw}, l_{prop}, and l_{AS} are the UE, radio access network (RAN), switching, propagation, and AS latency components, respectively [3]. The end-to-end service availability (a_{e2e}) is represented by Eq. (2) as the product of the availability values of UE, RAN, TN, and AS availability, denoted as a_{UE}, a_{RAN}, a_{TN}, and a_{AS}, respectively [3]. The availability of a TN path (a_{TN}) is computed as in Eq. (3), where n is the number of traversed nodes, f_i^{node} is the failure probability of node i, f_j^{link} is the failure probability (per km) of link j, and d_j is the length of link j [3]. When protection is considered, two node-disjoint paths connect the AP to its AS.

The model used to compute the cost of *(i)* the TN resources and *(ii)* the computing infrastructure necessary to support the required service deployment, assumes that the RAN is already deployed, thus the model considers only the cost of the transceivers, the TN nodes, and the DCs hosting ASs. The transceiver cost depends on the supported data rate. The cost of the TN nodes varies with the node type. TN nodes that do not perform traffic aggregation are equipped with only optical cross connects (OXCs), multiplexers (MUXs), and de-multiplexers (DeMUXs). TN nodes performing traffic aggregation are also equipped with packet switches and transceivers. The total cost of the computing resources is determined by the required number of DCs, the number of ASs within each DC, and the cost scaling factor of a DC. This latter parameter is used to models how the intra-DC infrastructure, power, and networking costs are related to the cost of server [1]. For instance, when the cost scaling factor is 0.5, for each cost unit spent on servers, only 0.5 cost units will be spent on the infrastructure, power, and networking equipment. Large DCs have low cost scaling factors, conversely, small scale DCs present high costs for their infrastructure equipment.

3. Cost Assessment

This section assesses the trade-offs between the savings obtained by centralizing services on a few high-scale DCs and the extra cost required for protecting the TN paths while meeting the availability requirements of the deployed services. Four scenarios are considered (Table 1). In each scenario, the maximum AP-DC distance ($d_{CN} = \sum_{j=0}^{n+1} d_j$) is computed by taking into account the latency and availability constraints of a service and by solving the equations presented in Sec. 2, assuming links of the same length in the TN. The DC type, the cost scaling factor, and the number of APs that can be served by a single DC (i.e., service density) and the number of AS that can be hosted in a single DC are defined in Table 2. They vary as a function of the max AP-DC distance computed for a specific UC. The channel rate used in the local, province, regional, and national TN is 1, 10, 100, and 100 [Gbps], respectively. Each fiber carries 80 channels. The numerical results presented in this section are obtained using the models described in Sec. 2 considering 20,000 APs, each one requiring a rate of 1 [Gbps]. We assume that l_{RAN}=3 [ms], l_{sw}=0.2 [ms] for all the nodes, $f_j^{link} = 3 \times 10^{-5}$/km ($\forall j$), $f_i^{node} = 10^{-6}$ ($\forall i$), while $a_{UE} = a_{AP} = a_{AS} = 1$. For the TN cost model, we assume that the cost of an OXC and a packet switch increases linearly with the number of ports. The cost of a 100 [Gbps] transceiver is defined as τ times the cost of a 10 [Gbps] transceiver. The values used in the cost model are listed in Table 3.

Figure 2(a) shows the maximum AP-AS distance as a function of the number of TN nodes to be traversed, when a_{RAN}=0.99999. The figure shows that the maximum AP-AS distance that can be traversed over the TN can be greatly increased by adding redundant resources, i.e., the protected (P) case. For UCs 1 and 4 the maximum AP-AS distance can go slightly beyond 1000 [km], without violating their respective 12 [ms] and 20 [ms] latency constraints. For UCs 2 and 3, adding redundant resources to the TN paths allows the maximum distance to increase from a few tens to around 300 [km], while still being within the 5.5 [ms] and 20 [ms] latency constraint. Fig. 2(b) shows the cost savings obtained when high-scale DCs can be reached thanks to longer TN paths. The cost savings are a function of the γ parameter used in Table 2. Applications with relaxed latency and strict avail-

Table 1: UCs considered.

UC	Description	Target latency	Availability	Reference
1	Augmented Reality, collaborative gaming	12ms	99.9%	[3]
2	Remote control for smart manufacturing	5.5ms	99.99%	[3]
3	Discrete automation	20ms	99.99%	[4]
4	Process automation / Monitoring	20ms	99.9%	[4]

Table 2: DC characteristics.

Type	AP distance [km]	Service density	Num. AS	Cost scaling factor
National	d_{CN}>100	1000	250	N_{EF}=0.5
Regional	10<d_{CN}≤100	100	25	R_{EF}=$\gamma \times N_{EF}$
Province	1<d_{CN}≤10	10	3	P_{EF}=2 × R_{EF}
Local	d_{CN}≤1	2	1	L_{EF}=2 × P_{EF}

Table 3: Cost values.

Resource	Cost [CU]
MUX / DeMUX	1 [5]
1G/10G Transceiver	3.2 / 51.5 [6]
OXC port	0.5
1G/10G/100G, SW port	0.6 / 0.9 / 4.3 [7]
Application server	27.8 [8]

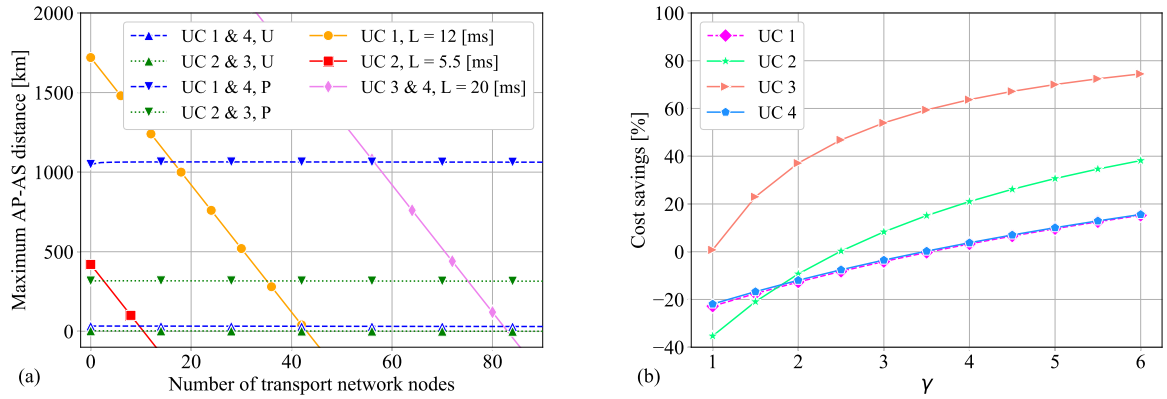

Fig. 2: (a) Maximum AP-AS distance for the unprotected (U) and protected (P) case with a_{RAN}=0.99999. (b) Cost savings by allowing redundant resources over the TN varying γ.

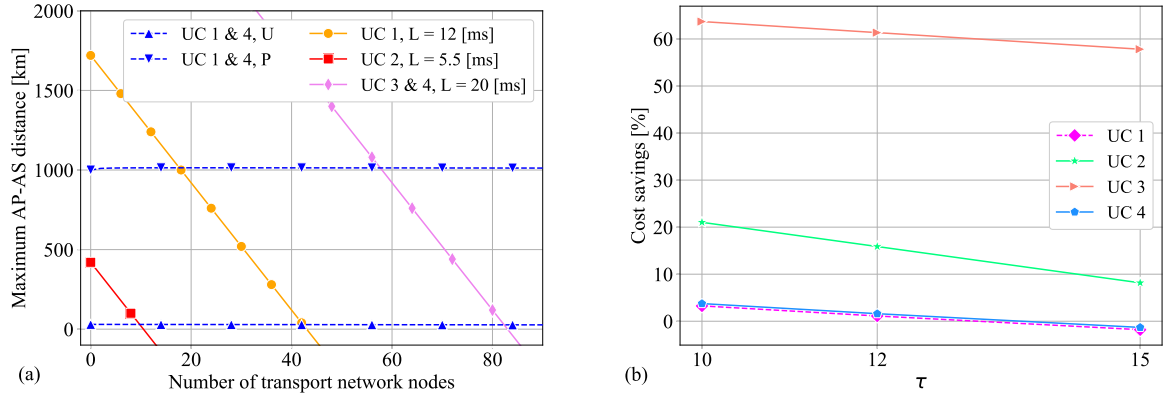

Fig. 3: (a) Maximum AP-AS distance for the unprotected (U) and protected (P) case with a_{RAN}=0.9999. (b) Cost savings by allowing redundant resources over the TN varying τ with γ=4 and a_{RAN}=0.99999.

ability constraints (i.e., UC 3) benefit the most with up to 74% cost savings when $\gamma = 6$. Figure 3(a) presents the same results as Fig. 2(a) but considering the case when a_{RAN}=0.9999. In this case, it is not possible to meet the availability constraints of UCs 2 and 3. Moreover, the maximum AP-AS distance for UC 1 for the protected case is slightly reduced compared to the one shown in Fig. 2. This is due to the reduced value of a_{RAN}. However, the cost savings are still significant. The values are not reported here because of space constraints but the pattern is similar to the one presented in Fig. 2(b). Finally, Fig. 3(b) shows how the cost savings vary as a function of τ, i.e., the ratio between the cost of a 10 and a 100 [Gbps] transceiver, when γ=4 and a_{RAN}=0.99999. UC 3 is the least affected by the variations of τ. This is because most of the cost savings come from the use of cost-effective computing resources rather than from the use of high rate channel in the TN. This is not the case for UCs 1, 2 and 4. In particular UC 2 is the most affected because most of the cost savings derive from being able to use 100 [Gbps] transceiver when redundant resources are added over the TN paths.

4. Conclusions

This work investigated the cost benefits of centralizing services on a few high-scale DCs. Despite the need to deploy extra (i.e., redundant) connectivity resources in the TN (i.e., to meet the availability requirements of some of the deployed services), results show that the economy of scale benefits from processing services in a few high-scale DC locations allows an up to 74% reduction of the overall infrastructure cost.

Acknowledgements: 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.

References

1. A. Greenberg *et al.*, "The cost of a cloud: Research problems in data center networks," CCR (2008). DOI: 10.1145/1496091.1496103.
2. P. Öhlén *et al.*, "Data plane and control architectures for 5G transport networks," JLT (2016). DOI: 10.1109/JLT.2016.2524209.
3. MGMN Alliance, "5G extreme requirements: End-to-end considerations," White paper (2019). Version 2.5.
4. 3GPP, "Service requirements for the 5G system; Stage 1," TS 22.261 (2019). Version 16.8.0.
5. "Fiberise 2-channel CWDM Mux/Demux module," Accessed: 2019-08-12.
6. "Huawei SFP-GE-LX-SM1310 and SFP-10G-ER-1310," Accessed: 2019-08-12.
7. "Huawei S5720-52X-SI-48S, FS S5850-48S6Q and FS N8500-32C," Accessed: 2019-08-12.
8. "Dell PowerEdge R740 rack server," Accessed: 2019-08-12.