Reducing operational costs of ultra-reliable low latency services in 5G

József Varga, Attila Hilt, József Bíró, Csaba Rotter, and Gábor Járó

***Abstract*— Ultra-reliable low latency (uRLL) communication in 5G dictates the deployment of distributed infrastructure with numerous datacenters for low latency, while hosting ultra-reliable services mandates attended datacenters. This would boost the operational costs of 5G network operators planning country-wide coverage for uRLL services. This paper examines how these operational expenses dominated by administrative costs can be reduced without impacting the quality of the provided uRLL service. Our results indicate that hosting uRLL services in unattended datacenters with increased hardware redundancy schemes can produce significant cost savings.**

***Index Terms*—G, availability, low latency, redundancy, total cost of ownership, cloud, datacenter**

I. INTRODUCTION

ULTRA-reliable low latency (uRLL) communication introduces stringent requirements for 5G systems [1], [2], [3]. A recent cost study [4] shows that provisioning uRLL services can represent significant part of complete 5G deployments. This is triggered by the demanding service requirements: (i) for ultra-reliability the standard solution is to deploy Tier-4 attended datacenters [5], [6]; (ii) for low latency these datacenters must be placed either at the edge or close to the edge of the core network.

3GPP has an ongoing study on enhancements for the support of uRLL communication [7]. This study only sets recommendations for the future normative specification work and assumes that the 5G system defined in [8] will be used as a baseline architecture for uRLL communication. Even though the detailed architecture specification for uRLL communication is still ongoing, for the deployments the general 5G system deployment assumptions are valid: it will utilize technologies like software defined networking and network function virtualization. Furthermore, due to the stringent latency requirements we can assume that application functions supporting uRLL services are not only hosted by the same infrastructure as the 5G core system but are co-located or even combined with 5G network functions.

In this paper, we examine how the costs of uRLL service deployments can be reduced. The paper is structured as follows: section 2 discusses how the latency requirements of uRLL services enforce the introduction of costly distributed deployment using coverage for Hungary as an example, and lists potential cost reduction options; section 3 deals with the possible unattended operation of datacenters hosting only low traffic volume of uRLL services [9]; section 4 presents the potential cost saving results compared to the costs of standard operation described in [4]; finally, section 5 concludes the paper.

J. Varga is with NOKIA, Budapest, Hungary. E-mail: jozsef.varga@nokia.com.
A. Hilt is with NOKIA, Budapest, Hungary. E-mail: attila.hilt@nokia.com.
J. Bíró is with NOKIA, Budapest, Hungary. E-mail: jozsef.biro@nokia.com.
Cs. Rotter is with NOKIA, Budapest, Hungary. E-mail: csaba.rotter@nokia.com.
G. Járó is with NOKIA, Budapest, Hungary. E-mail: gabor.jaro@nokia.com.

II. COSTS OF DEPLOYMENT FOR URLL SERVICES

The best-known example for uRLL services is the vehicle-to-everything (V2X) communication, including use cases like cooperative driving maneuvers (e.g. platooning), basic safety message, and see-through-system [10]. In densely populated countries the density of road system mandates nation-wide coverage. We use Hungary as an example to demonstrate how the uRLL service requirements are considered in planning the serving infrastructure.

II. 1 Infrastructure for uRLL Services

The end-to-end (E2E) latency requirements for V2X use cases available in the related literature and publications range from 3.3 millisecond (ms) [10], sub 10 ms [11], to 10-15 ms [2]. Independently of the actual value selected for E2E latency budget, it is further divided to elements like: service request processing at end user application, transmitting and receiving data at air interface, forwarding data in wired core network (fiber-optic), switching in packet data network, and optionally request handling in a server, see Fig.1. The use of these elements depends on what entities the uRLL service requires the network to connect: (i) a user to a server, (ii) two users using the network infrastructure, or (iii) users through a server.

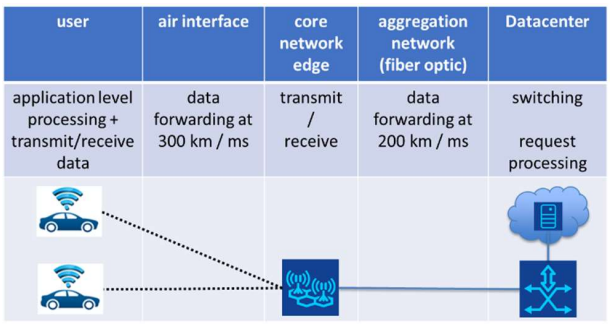

Fig. 1. Latency elements in 5G network.

DOI: 10.36244/ICJ.2018.4.6

The infrastructure serving the uRLL service is determined by the latency budget assigned for data forwarding in fiber-optic: as an example, if no time is assigned, then the uRLL service must be provided by mobile edge computing; while if 1 ms is assigned, then it allows data forwarding as far as 200 km in fiber-optic cable length. This "cable length budget" determines how the datacenters serving uRLL communications (for simplicity we use the term "low latency datacenter", LL DC in short) must be located to provide nationwide coverage. For the user-server-user scenario the E2E communication includes two "data forwarding in fiber-optic" elements and to avoid further splitting of the "cable length budget", we assume that for a specific uRLL communication session at a specific time all network functions are hosted in one LL DC, and as the server processing time is also strongly limited, we can also assume that all network functions are combined for uRLL communication. Note that the evaluation part of [7] suggests that servers needs to be kept geographically and topologically close to the user equipment, "within a transmission latency of 0.1 ms to 1 ms from the radio base station site".

Also, if the LL DCs are placed to serve the user-server-user scenario, then the resulting setup can appropriately serve the less demanding user-server scenario as well.

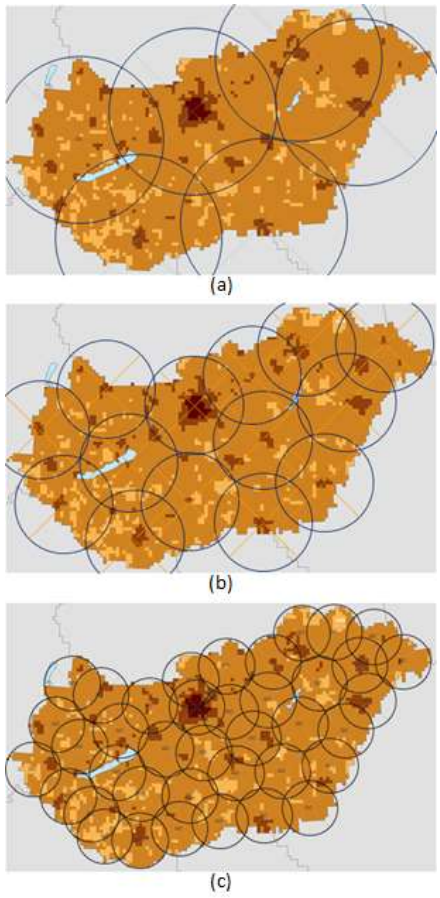

Fig. 2. LL DCs for Hungary with 100 / 60 / 33 km coverage radii.

The "cable length budget" determines the area covered by a LL DC, although for exact calculations the actual topology of the aggregation network is needed. For general calculations we estimate these areas as circles drawn around the potential locations of LL DCs. For this, we assume that the fiber-optic cable length connecting two points is usually 2-3 times longer than the geographical distance of those points.

As an example, Fig. 2 shows three options for LL DC placement in Hungary with different latency and cable length multiplier (mL) assumptions (copyright note: the population density map of Hungary is based on [12]).

Note that this circle-based coverage is only an estimate. In real-life deployments for uRLL services the locations of LL DCs must be designed based on the real cable lengths in the aggregation network.

In our model we did not aim to find coverage with minimal number of LL DCs, or full coverage, we just tried to place data centers to bigger cities wherever it was possible. Table 1 lists the estimated number of LL DCs needed to provide a country-wide coverage with three combinations of the latency and the cable length multiplier parameter values.

Option 2a assigns 2 ms latency from the E2E latency budget for data forwarding in fiber-optic (allowing 400 km cable length E2E) and mL = 2 as a multiplier for geographical distance to fiber-optic cable length. Keeping the user-server-user V2X use cases in mind, this translates to a 100 km coverage radius for LL DCs. Note that the 2 ms latency budget for forwarding data in fiber-optic is most probably too generous.

Option 2b takes 1.5 ms latency and mL = 2.5 length multiplier as parameters, resulting in a 60 km coverage radius. For comparison, cost calculations in [4] use a 72 km coverage radius.

Option 2c takes 1 ms latency for data forwarding in fiber-optic and mL = 3 as a geographical distance to optical cable length multiplier, resulting in a 33 km coverage radius for LL DCs.

Fig. 2 and Table 1 illustrate well that the number of LL DCs required to provide nation-wide coverage increases quickly as the latency requirements become harsher. The relation in our model coverage with circles around LL DCs is quite clear: the covered area is a quadratic function of the coverage radius. Even though with smaller coverage radius the nationwide coverage is achievable with less overlap among the coverage areas of LL DCs, and the coverage areas can follow better the country boundaries, still any decrease in the "cable length budget" will result a quadratic increase in the number of LL DCs for a country-wide coverage.

TABLE I
ESTIMATES FOR NATION-WIDE COVERAGE WITH LL DCS

Figure	2a	2b	2c
Latency budget for fiber-optic	2 ms	1.5 ms	1 ms
Cable length multiplier, m_L	2	2.5	3
Coverage radius for LL DCs	100 km	60 km	33 km
Number of LL DCs for nation-wide coverage	6	14	39

This statement is obviously valid for real deployments as well when the actual topology of the operator's aggregation network with real-life fiber-optic cable length is considered.

II. 2 Cost Considerations

Results in [4] indicate that IT admin costs could be the most significant cost contributor of the planned 5G infrastructure. IT admin costs are dominated by the 24/7 on-site support required for all LL DCs, see Fig. 3 that compares the monthly infrastructure costs of the main 5G use case groups: in addition to the uRLL services, the massive machine type communication – mMTC, evolved mobile broadband – eMBB, and ultra-dense high broadband service – uHBS. Note that those calculations use the standard cost calculations for datacenters [13], the IT admin costs are boosted by the fact that even for LL DCs covering rural areas with low traffic volumes (and thus hosting only tens of servers) the employment of 5 IT administrators is required: 40 working hours (minus vacations and sick leaves) per week to cover 168 hours a week.

In this work we focus on the IT admin cost reduction possibilities and leave other assumptions of [4] unchanged. For the 5G architecture it is assumed that (i) virtualization technologies [14] are used in all datacenters (including LL DCs), (ii) hard switches provide connectivity in the DC and may implement some services, e.g. user plane gateways, if they are SDN enabled [15], (iii) no other specialized hardware are deployed.

The obvious IT admin cost reduction possibility for LL DCs is to host additional services and share the IT admin costs with those services. Note that the intra-operator datacenter sharing possibilities are already considered in [4]. For example, datacenters hosting services like massive machine type communication and evolved mobile broadband on a national level, are also used as LL DC. In our example we can assume that the LL DC for the capital area is colocated with the datacenter hosting national level services, and thus the IT admin costs for uRLL services are already shared in that datacenter.

However, IT admin cost sharing may not be viable option for most LL DCs. As shown in our example, for option 2a it is possible to place most of LL DCs into bigger cities (population of 100,000+ in case of Hungary, obviously this depends on the population density and the level of urbanization of the country considered), but switching to the more realistic datacenter coverage options, for option 2b it is still possible to place all LL DCs into cities (again this statement is country dependent), but for the majority of the LL DCs these are already smaller cities and it is not expected to have significant demand for database capacities (i.e. no sharing). Finally, for option 2c approximately 80% of the LL DCs are placed at rural areas. Note that the ratio of LL DCs in rural areas are even worse for big countries, as the country level road length is significantly higher).

Another IT admin cost reduction option is to host the uRLL services in 3rd party datacenters that meet the ultrareliability requirements. However, this option has the same limitation as sharing. Furthermore, if the 3rd party datacenter is not close enough to the operator's aggregation network, the extra routing further limits the available time in the E2E latency budget.

The third option is the unattended operation of datacenters hosting uRLL services only. To ensure that the reliability requirements are still met, this must be compensated by deploying additional redundant hardware. In section 3 we examine the feasibility of this option.

III. UNATTENDED DATACENTERS FOR URLL SERVICES

We will examine how the lack of on-site IT support affects the service availability in LL DC and how it can be compensated by additional redundant hardware in LL DCs.

Obviously, high service reliability and high service availability are not equivalent terms. However, maintaining the same high service availability with high mean uptime values and keeping the same serviceability at the same time guarantee unchanged high reliability. For serviceability, the software in LL DCs is maintained remotely as virtualization technologies include centralized management and orchestration [14], while regular hardware maintenance can be provided without on-site IT administration as well.

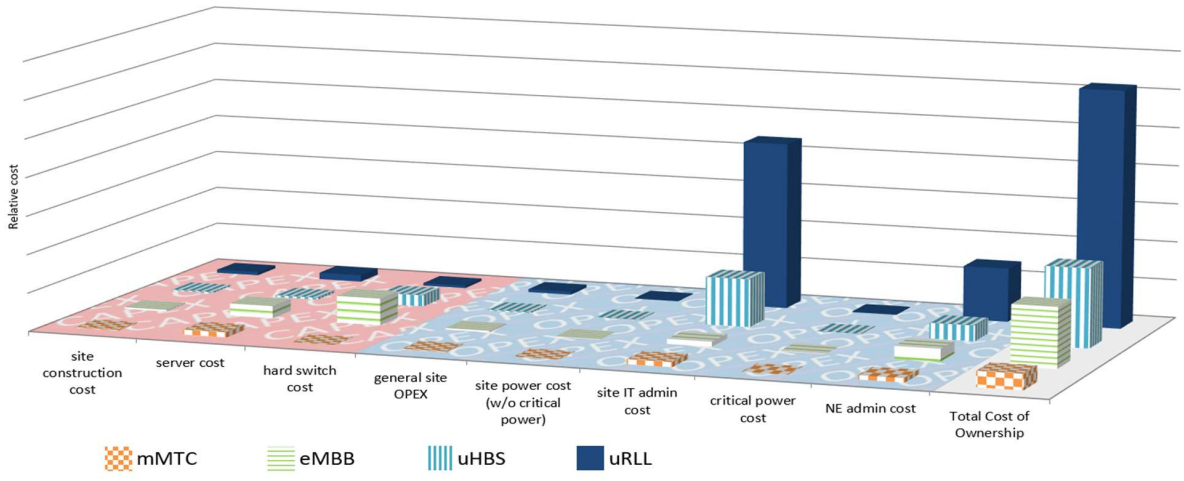

Fig. 3. Monthly total cost of ownership (TCO) of infrastructure for 5G use case groups [4].

III. 1 Assumptions

The assumptions used for LL DCs operated without on-site IT admin support are:

- those LL DC provide coverage for rural areas with low traffic volumes, thus the services are provided with minimal software configurations;
- those LL DC hosts only uRLL services;
- virtualization technologies [14] allow the remote management and configuration (automated and/or centralized) of software in those LL DCs;
- the virtual network functions (VNFs) are deployed on commodity servers;
- for hard switches and critical VNFs the 2N redundancy scheme is deployed;
- for other network functions utilizing load sharing redundancy scheme is the most efficient option [16];
- hard switches in LL DCs provide connectivity of the datacenter and may implement some services, e.g. user plane gateways, if they are SDN enabled [15];
- for best utilization of E2E latency budget the uRLL services are implemented in a single VNF (if any) and SDN based hard switches may implement user plane gateway functions.

III. 2 Considered Availability Parameters

The components of service availability in a datacenter are: software availability, hardware availability, datacenter infrastructure availability, and network availability. The dependency of these components with a high-level description is given in Fig. 4. The existence of on-site IT administration has impact only on the availability of IT hardware units by minimizing the downtime of hardware units (commodity servers and hard switches for the considered LL DCs). Note that regular maintenance impacting DC availability is still provided.

The hardware availability for network functions imple- menting uRLL services in LL DCs depends on both the availability of single hardware units (commodity server or hard switch, or even network interface cards) hosting the network function and the redundancy scheme applied for these hardware units. To obtain generic results we consider a range of typical values for uptime and downtime of a single hardware unit.

2N means full redundancy. For one single unit (N=1) 2N means duplication, such avoiding single point of failure. In case of several parallel units (N>1), all units are spared. In the system N units are required to support the traffic, but 2N units are deployed to increase the availability of the system. In 2N redundant systems, the system is capable to run as long as N units are available. 3N redundancy adds, for each unit carrying the load, two spare units in parallel. Please note, that the redundant units are not added to increase the system capacity: this way the system availability is improved. In 3N redundant systems, similarly to 2N redundancy, the system is capable to run as long as N units are available. "2 of 3" redundancy means that at least 2 units must operate from the total 3 units. Similarly, "2 of 4" means that minimum 2 units must operate from the total amount of 4 units. In the general case, N working units are spared by K redundant units, usually referred as N+K. The redundant units are either working or they are in standby mode. In active/active mode all the N+K units are working, and they share the total load (load-sharing). When K units are in standby mode, depending on the speed of launching the standby units into operation, we can talk about hot, warm or cold-standby. For the system performing its desired function, from the total number of N+K units at least N must operate, or in other words maximum K units can fail. The different sparing methods are discussed in detail in [17] and it is shown that in distributed systems load sharing efficiently increases the overall system availability.

For the redundancy schemes, we check all variants that match our assumptions (minimal configurations both for full redundancy and for load sharing based solutions). The examined values are:

- for the initial hardware redundancy scheme (i.e. when attended operation of LL DC is assumed) 1+1, 2+2 and 2+1 (also known as 2 of 3) redundancy;
- for uptime values, the mean time between failures (MTBF) parameter is used. The considered values are 200,000, 300,000, and 400,000 hours. Note that commodity server and hard switch vendors do not publish concrete values nowadays. Therefore, we adopted the typical values used by web pages and literature discussing availability, such as [18] and [19];
- for downtime values, the mean time to repair (MTTR) parameter is used. The considered values are 10, 20, 30, 40, 60, and 90 minutes. The widely used MTTR estimate for classical DC environment is 60 minutes, but as for 5G we expect high level of automation [20] including software management and configuration [14], our study focuses mostly on lower values.

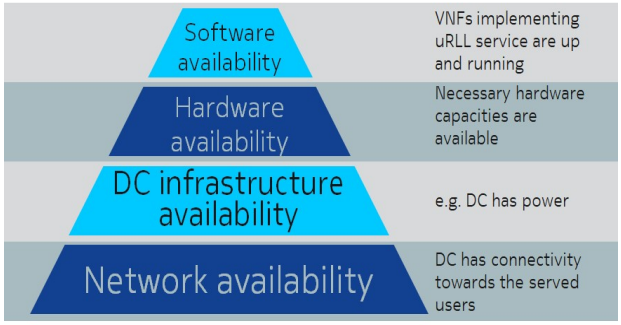

Fig. 4. Main components for service availability in a datacenter..

III. 3 Availability Calculatiuon Details

For all combinations of the above parameters we first calculate the hardware availabilities with base redundancy schemes, then calculate the hardware availabilities using a more stringent redundancy scheme – obviously, receiving higher hardware availability values. Finally, we start to increase the MTTR value until the hardware availability with the more stringent redundancy scheme and increased Fig. 5. Base and increased hardware availability for MTBF=400,000 hours and MTTR = 10 minutes initial parameters.

MTTR drops back to the same value as with the base redundancy scheme and the initial MTTR value. The increase

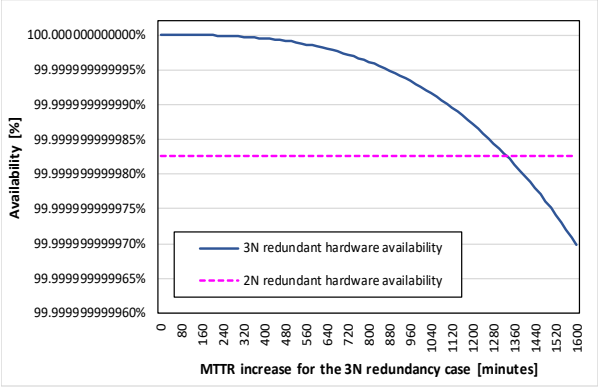

Fig. 5. Base and increased hardware availability for MTBF=400,000 hours and MTTR = 10 minutes initial parameters.

in MTTR provides the additional time for IT administrators to travel to LL DCs.

As an example, this calculation is illustrated in Fig. 5, with MTBF set to 400,000 hours, MTTR set to 10 minutes, and redundancy scheme for the network function implementation of the uRLL service is assumed to be 2N redundancy. The base hardware availability for this case is A2N=99.9999999999826 %. Switching to 3N redundancy the hardware availability increases to an even higher value of A3N=99.99999999999999999927662127444 %.

Note that the above hardware availability numbers represent 12 and 19 9's availability. Obviously, these numbers are just one component in the service availability, which is expected to remain in the typical range of telco services, i.e. 5 or 6 9's availability.

Continuing with the example, the hardware availability for the 3N redundant case drops to the value of the base 2N redundant setup when the MTTR is increased by 1330 minutes to 1340 minutes. That is if a uRLL service is imple- mented on a 2N redundant hardware in an attended LL DC (assuming 400,000 hours MTBF and 10 minutes MTTR), then implementing the same uRLL service on a 3N redundant hardware guarantees the same hardware availability even if MTTR is increased to 1340 minutes.

The formulas used for basic availability calculation of a single hardware unit and the hardware availability for implementations are listed here for information, the details are available in [4].

The generic availability is defined in (1), MUT and MDT representing the mean uptime and mean downtime of the system, respectively.

$$Availability = MUT/(MUT+MDT) \quad (1)$$

The availability of a single hardware with the wellknown function based on MTBF and MTTR, as specified in (2). Note that other definitions also exist in the literature [16], [19].

$$A_{single}(MTBF, MTTR) = MTBF/(MTBF+MTTR) \quad (2)$$

The hardware availability for the 2N redundancy scheme is calculated as in (3), while for the 3N redundancy scheme is calculated according to (4).

$$A_{2N}(x, y) = 1-(1-A_{single}(x, y))^2 \quad (3)$$

$$A_{3N}(x, y) = 1-(1-A_{single}(x, y))^3 \quad (4)$$

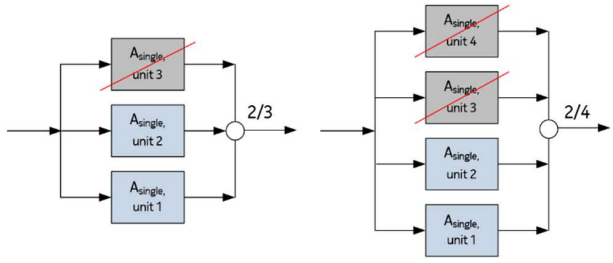

Fig. 6. "2 of 3" and "2 of 4" redundant parallel systems.

Hardware availability for the "2 of 3" redundancy scheme is calculated according to (5), while for the "2 of 4" redundancy scheme it is calculated according to (6).

$$A_{2/3}(x, y) = 3A_{single}(x, y)2-2A_{single}(x, y)^3 \quad (5)$$

$$A_{2/4}(x, y) = 1-(4(1-A_{single}(x, y))3-3(1-A_{single}(x, y))^4) \quad (6)$$

The "2 of 3" and "2 of 4" redundancy schemes are depicted in Fig. 6. These schemes are regularly referred as "2+1" and "2+2" as well.

III. 4 Results

Previously we have explained our calculations for a specific parameter set, here the results for all parameter combinations are summarized:

- single hardware availability values for all (MTBF, MTTR) combinations are in Table 2;
- hardware availability values both for 2N and 2+1 (or "2 of 3") redundancy schemes are well above the usual service availability values (more than ten 9's), so for better illustration we include the unavailability values for these cases: the unavailability values in Table 3 are for

TABLE 2
SINGLE HARDWARE UNIT AVAILABILITIES FOR THE EXAMINED MTBF, MTTR RANGE

A_{single}(MTBF, MTTR) (as %)			
MTTR (minutes)	MTBF (hours)		
	200,000	300,000	400,000
10	99.999917	99.999944	99.999958
20	99.999833	99.999889	99.999917
30	99.999750	99.999833	99.999875
40	99.999667	99.999778	99.999833
60	99.999500	99.999667	99.999750
90	99.999250	99.999500	99.999625

TABLE 3
HARDWARE UNAVAILABILITY WITH 2N REDUNDANCY SCHEME

1-A_{2N}(MTBF, MTTR)			
MTTR (minutes)	MTBF (hours)		
	200,000	300,000	400,000
10	6.9444E-13	3.0864E-13	1.7364E-13
20	2.7778E-12	1.2346E-12	6.9444E-13
30	6.2500E-12	2.7778E-12	1.5625E-12
40	1.1111E-11	4.9383E-12	2.7778E-12
60	2.5000E-11	1.1111E-11	6.2500E-12
90	5.6249E-11	2.5000E-11	1.4062E-11

TABLE 4
HARDWARE UNAVAILABILITY WITH 2+1 REDUNDANCY SCHEME

$1\text{-}A_{2/3}$(MTBF, MTTR)			
MTTR (minutes)	MTBF (hours)		
	200,000	300,000	400,000
10	2.0833E-12	9.2593E-13	5.2081E-13
20	8.3333E-12	3.7037E-12	2.0833E-12
30	1.8750E-11	8.3333E-12	4.6875E-12
40	3.3333E-11	1.4815E-11	8.3333E-12
60	7.4999E-11	3.3333E-11	1.8750E-11
90	1.6875E-10	7.4999E-11	4.2187E-11

2N redundancy scheme, while the unavailability values in Table 4 are for 2+1 (or "2 of 3") redundancy scheme.

- The hardware availability values both for 2N and 2+1 (or "2 of 3") redundancy schemes are illustrated in Fig. 7.

The hardware availability values for the improved redundancy schemes are not illustrated in any tables or figures, as with the base MTTR the hardware availability values are "too close" to 1 (see the example earlier with 19 9's).

- Table 5 lists the potential MTTR increase (for all the considered MTBF/MTTR combinations) for LL DCs where uRLL services are originally provided by a single network function implemented with 2N redundancy, and the network function implementation is switched to 3N redundancy to allow unattended LL DC operation;
- Table 6 is similar, but the original network function redundancy scheme is 2+1, and it is switched to 2+2 scheme;
- Table 7 combines the two cases assuming that the uRLL service are originally provided by two network functions: an SDN based 2N redundant hard switch as a user plane gateway, and 2+1 redundant control plane network function.

Note that all the values for "MTTR increase" are rounded off to 10 minutes. For example, in Table 5 the value 1200 in the cell of MTTR 10 minutes and MTBF 300 k hours means (7). That is assuming 300,000 MTBF, if the MTTR is increased by

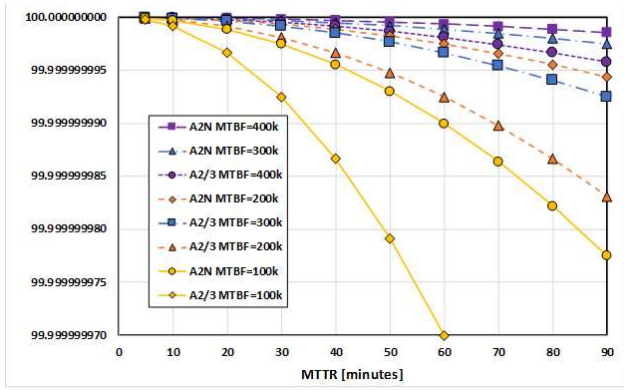

Fig. 7. Availability of 2N and "2 of 3" redundant systems.

TABLE 5
MTTR INCREASE COMPENSATED BY SWITCHING FROM 2N TO 3N REDUNDANCY SCHEME

compensated MTTR increase (minutes)			
MTTR (minutes)	MTBF (hours)		
	200,000	300,000	400,000
10	1050	1200	1330
20	1660	1910	2100
30	2180	2500	2750
40	2630	3020	3330
60	3440	3950	4360
90	4500	5170	5700

TABLE 6
MTTR INCREASE COMPENSATED BY SWITCHING FROM 2+1 TO 2+2 REDUNDANCY SCHEME

compensated MTTR increase (minutes)			
MTTR (minutes)	MTBF (hours)		
	200,000	300,000	400,000
10	950	1090	1200
20	1510	1730	1910
30	1970	2260	2500
40	2390	2740	3020
60	3120	3590	3950
90	4080	4690	5170

TABLE 7
MTTR INCREASE COMPENSATED FOR THE COMBINED CASE

compensated MTTR increase (minutes)			
MTTR (minutes)	MTBF (hours)		
	200,000	300,000	400,000
10	970	1110	1230
20	1540	1770	1950
30	2020	2310	2550
40	2440	2800	3090
60	3190	3660	4040
90	4180	4790	5280

1200 minutes to 1210 minutes, the 3N hardware redundancy scheme still provides the same hardware availability as the 2N hardware redundancy scheme with a 10 minutes MTTR.

$$A_{3N}(300000, 1210) > A_{2N}(300000, 10) > A_{3N}(300000, 1220) \quad (7)$$

The results show that the unattended operation for a LL DC that provides uRLL services by hosting at most a hard switch (for data forwarding) and an additional virtual network function for each uRLL service is possible. Furthermore, as the potential MTTR increase allowed by switching to more stringent redundancy schemes is practically above 16 hours for any practical combination of (MTBF, MTTR) value pairs, the IT administrators responsible for the LL DCs can be employed in a single shift. More details on the cost consequences are given in the next section.

IV. COST CONSEQUENCES

From the considered IT admin cost reduction possibilities, the hosting of 3rd party services in LL DC has a straightforward impact on the expenses of the uRLL services: the IT admin costs are reduced. We do not consider changes in other elements of the TCO: even though hosting 3rd party services requires deploying additional hardware capacities in LL DC means increase in both the capital expenses and the power consumption of the LL DC, it is fair to assume that the income of hosting activity compensates these additional costs.

As the dominating cost for uRLL services the IT admin costs are boosted by the fact that operators must employ 5 IT administrators for 24/7 supervision of relatively low number of servers, while normally an IT administrator may supervise hundreds of servers [13]. In [4] 500 servers per IT administrator are used in calculations. Hosting 3rd party services allows better utilization of the IT administrators. Note that even though operators may not share IT admin costs proportionally to the used hardware capacities (as an example if a LL DC requires 50 servers and hard switches to provide uRLL services, and an additional 100 hardware elements to host 3rd party services, the 3rd party tenants may not be ready to pay 2/3 of the IT admin costs, but rather a 20% (100/500) only), sharing still can provide significant IT admin cost save. Obviously to avoid cost increase instead of cost save, operators must find long term reliable tenants.

As shown in section 3, unattended operation of LL DC is also an appealing option. It changes the LL DC TCO as follows:

- due to the increased redundancy scheme applied, the capital expenses are increased by 50%. This includes site construction costs (which is calculated from the power consumption of the servers and hard switches deployed in the datacenter), server costs, and hard switch costs;
- critical power (server and hard switch power consumption) cost increased by 50%;
- site power consumption (excluding critical power) assumed to be unaffected (usually calculated from critical power consumption, but the impact of increased critical power consumption is compensated by the removed on-site IT administrators;
- on-site IT admin cost is replaced by the cost of centralized "IT administrator pool" assigned to the regular maintenance of rural LL DCs and visiting a LL DC whenever a hardware error is detected (software administration is centralized). This allows the reduction of 5 IT administrators per LL DC to 0.5 IT administrator per LL DC. But the IT administrator employment costs are increased by 50% (e.g. regular travels, company cars).

Note that it is also possible to outsource IT administrators, and it may result in further cost reduction for the operation of unattended LL DCs, but it is not in the scope of this study.

Combined application of the two studied concepts is also possible (i.e. hosting 3rd party services in unattended LL DC), but increased hardware capacity in the LL DC may require more frequent IT administrator visits, and our availability calculations assume simple uRLL service deployments only. Availability calculations for more complex service deployments are not considered here.

Fig. 8 illustrates the cost saving potential of shared LL DC and unattended LL DC concepts. The calculations are based on the results of [4] (the standard DC cost model) and the two studied concepts are applied with different weights: the "no sharing – 90% unattended" concepts intends to model a strongly urbanized large country (10% of the LL DCs are located in big cities hosting regional level DCs anyway, the rest of LL DCs practically cover the road system of rural areas). While the other figures intend to model densely populated countries either with smaller number of huge cities, or relatively high number of bigger cities.

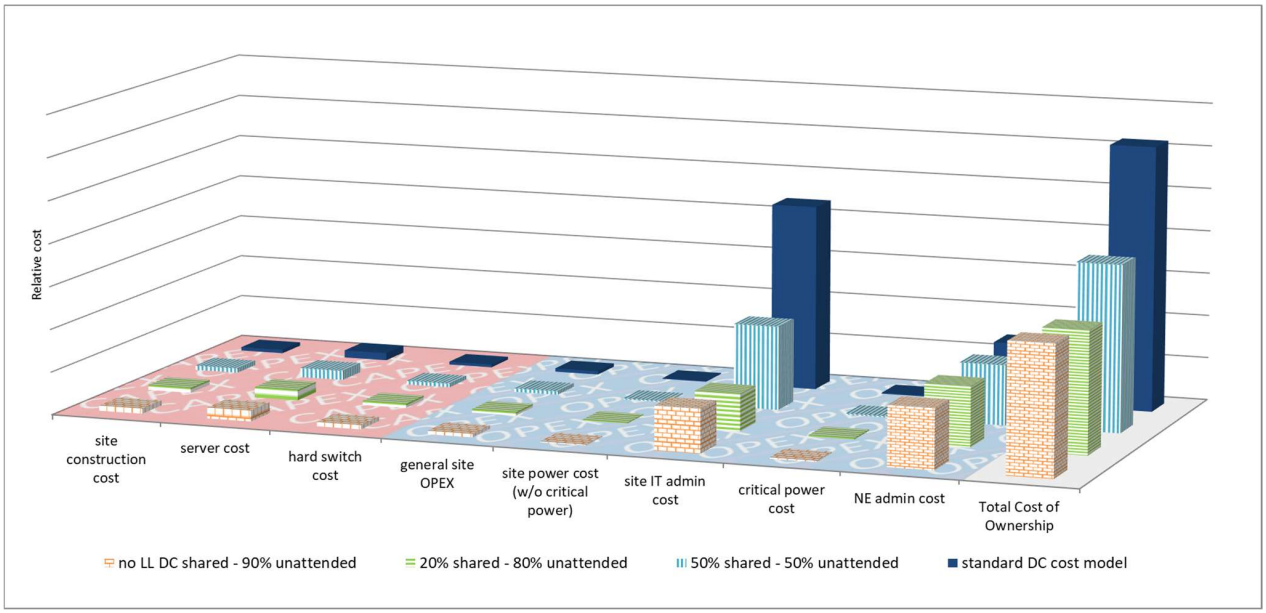

Fig. 8. TCO of uRLL services with different variations of LL DC operation concepts.

As cost calculations show, the cost saving potential is significant for both concepts. The applicability depends on the population density, urbanization, size of covered area, total road length of the operator's country and obviously the actual topology of the operator's aggregation network.

V. CONCLUSION

Initial study in 5G infrastructure TCO [4] made it obvious that for uRLL services standard datacenter operational model is way too expensive.

Both sharing LL DCs and unattended operation of LL DCs provide significant cost saving. The possibilities of sharing for LL DC can be limited geographically (big cities only), but it is nicely complemented by the unattended operation of LL DCs, which is applicable for rural areas with low traffic volumes.

The main drawback for unattended operation of LL DCs is being environmentally "unfriendly": with more servers and switches deployed the power consumption increases and this requires regular travel for visiting IT administrators. The latter can be reduced (potentially resulting further cost reduction) if operators outsource IT admin tasks.

Our current work focuses on maintaining service availability in a single LL DC, service availability for special cases involving multiple LL DCs, e.g. mobility related handover scenarios or services implemented by distributed applications, requires further study.

ACKNOWLEDGMENT

This work was supported by the National Research, Development and Innovation Fund of Hungary in the frame of FIEK_16-1-2016-0007 (Higher Education and Industrial Cooperation Center) project.

Thanks to Noemi Wagner for proofreading the text and finalizing figures and tables. The authors acknowledge Csaba Vulkán and Adorján Korényi, both with Nokia Hungary, for their continuous support.

REFERENCES

[1] NGMN alliance, NGMN 5G White Paper, 2015-02.

[2] 3GPP, TR 22.862; Feasibility Study on New Services and Markets Technology Enablers for Critical Communications; V14.1.0, 2016-09.

[3] 3GPP, TR 22.891; Feasibility Study on New Services and Markets Technology Enablers; V14.2.0, 2016-09.

[4] Kiess, W., Sama, M. R., Varga, J., Prade, J., Morper, H. and Hoffmann, K., "5G via Evolved Packet Core Slices: Costs and Technology of Early Deployments", *Proc. IEEE Symp. Personal, Indoor, and Mobile Radio Communications (PIMRC '17)*, pp. 1-7, Oct. 2017, DOI: 10.1109/PIMRC.2017.8292691.

[5] Varga, J., Hilt, A., Rotter, Cs. and Járó, G., "Providing Ultra-Reliable Low Latency Services for 5G with Unattended Datacenters", *Proc. IEEE Symp. Communication Systems, Networks & Digital Signal Processing (CSNDSP '18)*, pp. 1-4, Jul. 2018, DOI: 10.1109/CSNDSP.2018.8471756.

[6] Turner IV, W. P., Seader, J. H., Renaud, V. and Brill, K. G., Tier Classifications Define Site Infrastructure Performance, The Uptime Institute white paper, 2006.

[7] 3GPP, TR 23.725; Study on enhancement of Ultra-Reliable Low-Latency Communication (URLLC) support in the 5G Core network; V16.0.0, 2018-12.

[8] 3GPP, TS 23.501; System Architecture for the 5G System; V15.4.0, 2018-12.

[9] TIA-942-A, Telecommunications Infrastructure Standard for Data Centers, 2012.

[10] Alieiev, R., Kwoczek, A. and Hehn, T., "Automotive Requirements for Future Mobile Networks", *Proc. IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM '15)*, pp. 1-4, Apr. 2015, DOI: 10.1109/ICMIM.2015.7117947.

[11] Simsek, M., Aijaz, A., Dohler, M., Sachs, J. and Fettweis, G., " 5G-Enabled Tactile Internet", IEEE Journal on Selected Areas in Communications, vol. 34, pp. 460–473, March 2016.

[12] Gridded Population of the World," Hungary: Populationdensity, 2000", available at http://sedac.ciesin.columbia.edu/gpw.

[13] Barroso, L., Clidaras, J. and Hölzle, U., "The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines", Morgan & Claypool, 2nd edition, 2013.

[14] ETSI, GS NFV-MAN 001; Network Functions Virtualisation (NFV); Management and Orchestration; V1.1.1, 2014-12.

[15] An, X., Kiess, W., Varga, J., Prade, J., Morper, H. and Hoffmann, K., "SDN-based vs. software-only EPC gateways: A cost analysis", *Proc. IEEE NetSoft Conference and Workshops (NetSoft '16)*, pp. 146-150, June 2016, DOI: 10.1109/NETSOFT.2016.7502461.

[16] Hilt, A., Járó, G. and Bakos, I., "Availability Prediction of Telecommunication Application Servers Deployed on Cloud", *Periodica Polytechnica Electrical Engineering and Computer Science*, 60(1) pp. 72-81, 2016.

[17] Nagy, L., Hilt, A., Járó, G., Oláh, D. and Talyigás, Zs., "Comparison of Availability Figures of Distributed Systems Using Multiple Redundancy Methods", AACS'2016 Workshop, June 2016.

[18] Taylor, Z. and Ranganathan, S., "Designing High Availability Systems: DFSS and Classical Reliability Techniques with Practical Real-Life Examples", Wiley - IEEE Press, November 2013.

[19] Bauer, E., Adams, R., "Service Quality of Cloud-Based Applications", Wiley, IEEE Press, ISBN 978-1-118-76329-2, 2017.

[20] ONAP, Beijing Release Notes, V2.0.0, June 2018.

József Varga, PhD, is a senior research engineer at Nokia, member of the 'Multi Cloud Orchestration' research group in Nokia Bell Labs. He received his MSc in computer science and mathematics from the University of Szeged in 1991, PhD in IT from the University of Veszprém in 2002. He was an assistant professor at the University of Szeged, then at the University of Veszprém. He joined Nokia in 1999, he was involved in IP Multimedia Subsystem development, then represented Nokia as a standardization delegate in 3GPP from 2004 to 2011. In 2011 he joined Nokia Research Center (now Nokia Bell Labs) dealing with topics like SDN, virtualization, and orchestration. Currently he is focusing on resource management in 5G, including the economic aspects. He co-authored more than 10 papers and more than 10 granted patents.

Attila Hilt, PhD, graduated in Electrical Engineering from the Budapest University of Technology and Economics (BME) in 1990. In 1989 he joined the Research Institute for Telecommunications (TKI) Budapest, Hungary. Until 1999, he headed the Telecommunication Test Laboratory of TKI. In May 1999, he received the 'Doctor of INPG' degree from the Institute National Polytechnique de Grenoble, France. He received the PhD degree in 2000 from BME. He joined Nokia Networks in 2000 where he is currently working as senior system specialist. He has been involved in the dimensioning and design of several mobile and cloud networks. He participated in the testing, piloting, deployment and optimization of mobile and backhaul networks, including nationwide modernization projects in Europe. His research interests include measurement techniques, fixed, wireless, mobile and cloud-based communication networks, microwave, millimeter-wave and photonic systems. He authored/co-authored more than 80 papers and more than 120 research and project reports, guidelines and specifications. He is a member of the Scientific Association of Infocommunications (HTE) and the Hungarian Chamber of Engineers (MMK). Since 2016 he is a member of the Telecommunications Public Body of the Hungarian Academy of Sciences and since 2018 a member of the Telecommunications Scientific Committee (MTA TTB).

József Bíró is a senior research engineer at Nokia Bell Labs. He received his MSc in electrical engineering from the Budapest University of Technology and Economics (BME) in 1990, with specialization in Measurement and Industrial Control Technologies. He started his professional career at the Hungarian Research Institute for Particle and Nuclear Physics where he designed embedded software systems for scientific spacecrafts. He joined Nokia Research Center in 1998, where he first specialized in Java technologies (JVM in particular) and helped the adoption of Java for mobile platforms. From 2004 onwards, he worked on high availability technologies for telecommunication platforms. His related activities included industrial specification work in Service Availability Forum. Starting from 2010, he turned towards virtualization and cloud technologies, particularly dealing with the deployment and management of complex applications in the cloud. During this period, he participated in the EU FP7 project 4CaaSt, where he was a task leader within a work package. He is currently the member of Nokia Bell Labs 'Multi Cloud Orchestration' research group and his current focus is on advanced end-to-end orchestration for telco cloud platforms, NFV and beyond, with strong outlook towards 5G systems.

Csaba Rotter is heading 'Multi Cloud Orchestration' research group in Nokia Bell Labs. He is involved in different cloud related research topics targeting cloud native network service operation challenges in highly distributed multivendor environment. His special interest is on how to ensure performance related service level agreements for concurrent applications sharing the resources in the same distributed environment. His passion in automation started years before, when he was responsible for test automation concept development in large telecommunication systems. He received MSc in applied electronics at the Technical University of Oradea in 1995 and MSc in IT management at the Central European University Budapest in 2008. He joined Nokia in 1999, later Nokia Research Center (currently Nokia Bell Labs) in 2008.

Gábor Járó, PhD, is a chief architect at Nokia. He received MSc degree in Electrical Engineering at the Budapest University of Technology and Economics (BME) in 1994. In the same year, he joined the Department of the Micro-wave Telecommunication at BME, where he started working on his PhD degree. His research interest focused on noise in high-speed optical receiver and optical system, millimeter-wave signal generation in optical systems. He received the PhD degree in 2001 from BME. He joined Nokia in 1999, he has been involved in the design of several mobile and cloud network elements and networks. His research interests include measurements in mobile and cloud-based communication networks, design and realization of ultra-reliable low latency network elements. He authored/co-authored more than 50 papers and several granted patents.