Published in: 2022 IEEE Future Networks World Forum (FNWF)
DOI: 10.1109/FNWF55208.2022.00120

The Cost of Uncertainty: Impact of Overprovisioning on the Dimensioning of Machine Learning-based Network Slicing

Caner Bektas, Stefan Böcker, Christian Wietfeld
Communication Networks Institute, TU Dortmund University, Otto-Hahn-Strasse 6, 44227 Dortmund
Email: {caner.bektas, stefan.boecker, christian.wietfeld}@tu-dortmund.de

***Abstract*—Increasing automation of industry verticals and frequently changing production cycles require a high level of production line modularity and are locally accompanied by frequently changing disjunctive application requirements. Thus, current and future wireless communication networks need to face the challenge of providing opportunities to rapidly adapt the network to its changing application demands in order to guarantee a resilient and interference-free communication. A possible key technology for implementing such a solution is represented by private 5G networks that are additionally equipped with network slicing in order to be able to meet the versatile requirements of novel applications. However, resilient network design as well as network slice dimensioning can only be guaranteed through detailed network planning. This requires expert knowledge, which is not yet present at most companies or institutions. Accordingly, automation of the network planning process is a possible solution. Existing coverage planning frameworks are extended by capacity planning in this work, and network slicing is introduced. It is shown on the basis of a realistic scenario that the predictability of data (e.g., traffic characteristics in low-latency slices) significantly influences capacity planning and must be taken into account in the dimensioning of 5G and beyond future mobile networks.**

I. Introduction

With the introduction of spectrum usage permits in various countries such as the USA, Germany, and others, there is increasing interest in private mobile networks based on 5G and beyond. Private 5G, especially in combination with network slicing, is seen as a key technology for industrial users who are increasingly focusing on digitization and automation of their processes [1], with first measurement studies evaluating the performance of this approach [2]. However, to ensure the safe operation of such private 5G networks, network planning is required. When applying for frequency use, the private operator of the network must ensure that interference with neighboring cells is mitigated by complying with specified limits (e.g., received power) at the cell edges. However, it cannot be assumed that the required expert knowledge about network planning exists at applicant companies or institutions. Thus, it makes sense to automate the entire network planning process. The authors have already developed an automated framework based on unsupervised learning in the context of coverage planning [3]. In this paper, an extension of this framework to capacity planning is presented, which also integrates new aspects of network planning that will be important in 5G and beyond. Different facets of this are detailed in the next section.

A. The Problem of Data Uncertainty and Safety Margins

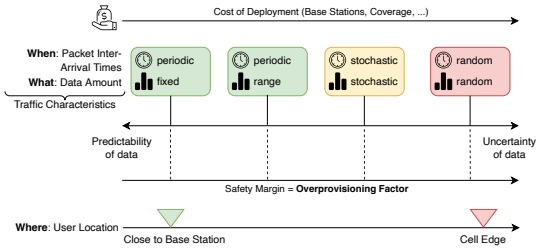

Fig. 1. Data uncertainty versus safety margins (or overprovisioning) of communication network resources [4].

To understand why novel capacity planning methods are needed for the dimensioning of future networks, Fig. 1 shows the relationship between data uncertainty and safety margins. The trend of deployment costs is shown above, with different network, participant, or application parameters varying significantly, depending on specific traffic characteristics. These parameters can be divided up into packet Inter-Arrival Times (IATs), data amounts, and user locations. For instance, in order to maintain zero scheduling latency in critical Ultra-Reliable Low Latency Communications (uRLLC) slices, the IATs and data amounts of the network slice participants have to be predicted to avoid lengthy scheduling request and grant mechanisms. This means that the cost of deployment, e.g., amount of base stations to deployed, rises with the uncertainty of data. For example, periodic packets with a fixed data amount are easily predictable, even without Machine Learning (ML), whereas stochastic processes are harder to predict perfectly. The most extreme case of uncertainty arises with random, rare events, where it is nearly impossible to make a forecast for the network scheduler. With rising uncertainty, the network operator has to provide safety margins, or overprovisioning, in order to mitigate prediction errors, which in turn raises the costs of deployment. Also, the location of the users determine the signal quality, varying the resources needed for upcoming data transfers. Different trade-off strategies can be considered in order to balance out these relationships, e.g., static assignment of resources, where unused resources are wasted, or by deploying more base stations to increase the average signal quality.

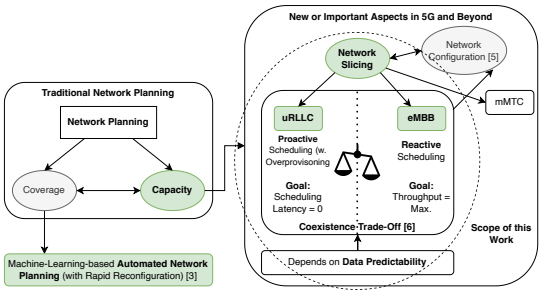

Fig. 2. Extension of traditional network planning to meet the challenges of 5G and beyond networks.

This means that traditional network planning aspects have to be extended in order to integrate these new aspects. In Fig. 2 left, the aspects of traditional network planning are depicted, which are coverage and capacity planning. The details of coverage planning and methods for its automation were presented by the authors in a previous work [3]. As for capacity planning (cf. Fig. 2 right), new or important aspects are emerging especially for 5G and beyond communication networks. The main aspects are, among others, the importance of network configuration [5] as well as network slicing for the safe operation of certain application and service types, especially for industrial verticals [6]. Typically, three main service types are identified for network slicing [7]: *uRLLC* for very low latency and high reliability (e.g., remote control of vehicles or robots), *Enhanced Mobile Broadband (eMBB)* for very high throughput and high spectrum efficiency (e.g., video / live event streaming), and *Massive Machine Type Communications (mMTC)* focusing on very high connection density and network energy efficiency (e.g., smart city). In this work, the balance between uRLLC and eMBB slices is analyzed in the context described in the introduction. The important aspect here is the coexistence trade-off between these two service types which depends on data predictability.

B. Importance of Data Predictability for Capacity Planning

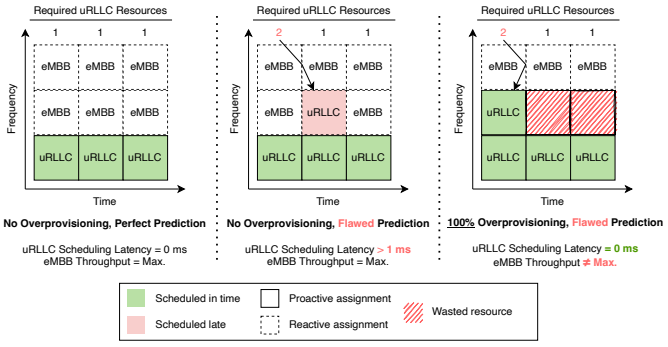

Fig. 3. The effects of prediction quality (predictability), scheduling mechanisms, and overprovisioning on uRLLC latency and eMBB throughput [6].

In Fig. 2 right, the main differences between eMBB and uRLLC slices are described. Regarding the scheduling of data packets, eMBB slices can safely be scheduled traditionally with reactive and latency-inducing scheduling request and grant mechanisms, as the main goal is to maximize the data throughput and scheduling latency can be tolerated. On the contrary, the uRLLC slice data traffic requires very low latency and thus, the elimination of any possible components of the end-to-end latency. In [6] the authors showed that the scheduling latency of uRLLC slices can be eliminated by predicting the packet IATs and data amounts of its User Equipments (UEs). However, the prediction quality is highly dependent on the *predictability* or *uncertainty* of the data and the limitations of utilized ML models. Fig. 3 illustrates this relationship by depicting various situations, where data predictability is varied. The three situations show a scheduling grid in the uplink comprising the time (x-axis) and the frequency (y-axis) domains. For illustrative purposes, three Resource Blocks (RBs) per Transmission Time Interval (TTI) can be scheduled, which are distributed by the scheduler to the uRLLC or eMBB slice. The TTI is $1\,\mathrm{ms}$ or lower, depending on the Subcarrier Spacing (SCS) [7]. Above, the required uRLLC resources are depicted, e.g., 1 meaning that a single RB is sufficient. On the left side of Fig. 3, the first situation is shown where all the required resources for the uRLLC slice can be scheduled since the ML-based scheduler was able to perfectly predict the packet IATs and the required RB. Thus, there is no scheduling latency in this case, since the packets can be sent immediately by the UEs as soon as they are queued in their own buffers. In addition, the maximum possible throughput can be achieved in the eMBB slice, since the uRLLC slice (with higher priority) only uses as many resources as required and thus the remaining capacity can be used for the throughput of the eMBB slice. In Fig. 2 middle, the number of RB required for the uRLLC slice increases in the first TTI. However, since the prediction fails due to poor predictability, only one RB is granted in advance even though two were needed. Thus, the packet must be held back at least one TTI (in most cases much more) in the UE's buffer to be sent in the next TTI, or even later after a lengthy scheduling request and grant sequence. Depending on the SCS and thus the Scheduling Request (SR) periodicity, the resulting latency is $>1\,\mathrm{ms}$ [7], which is not optimal for low-latency slices. However, by not wasting any resources, the maximum throughput in the eMBB slice is preserved. In Fig. 3 on the right is the same initial situation as above, but here an overprovisioning is configured. Since this is a $100\,\%$ overprovisioning, twice as many resources are reserved for the uRLLC slice as they should have been according to the prediction. Thus, incorrect predictions can be intercepted, as is done in this case, to keep the scheduling latency in the uRLLC slice at $0\,\mathrm{ms}$. However, it is obvious that in the case of a correct prediction the previously allocated resources are wasted by overprovisioning, and thus cannot be used in the eMBB slice. Thus, the actual maximum throughput can no longer be utilized, reducing the spectral efficiency.

The described relationship leads to the fact that in capacity planning, the uncertainty of data in low-latency slices must be taken into account. In this work, this is investigated in the following sections based on a realistic scenario.

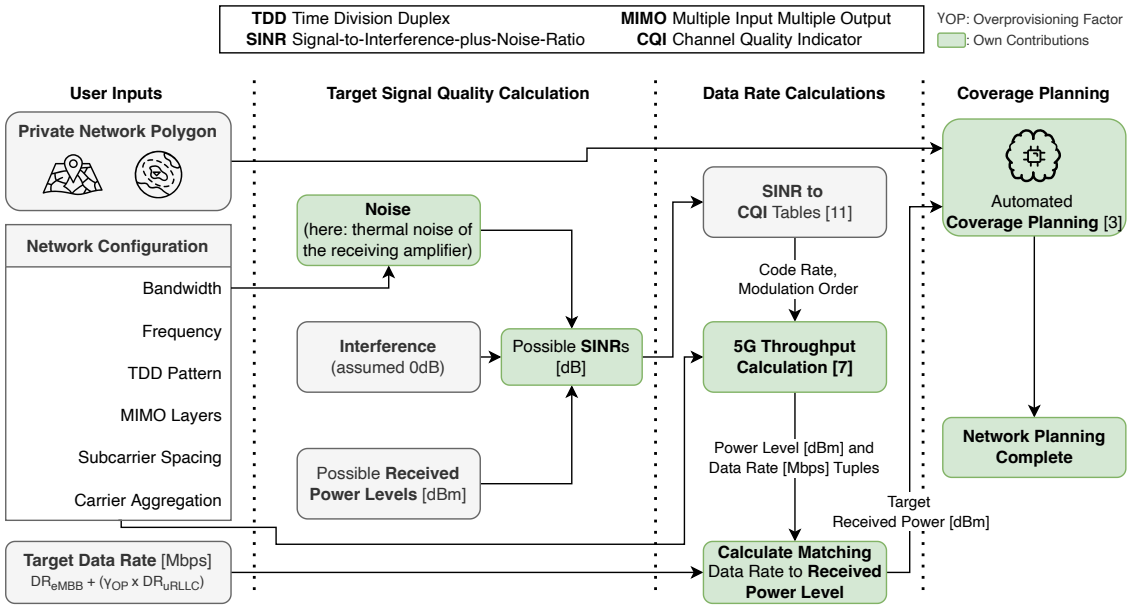

Fig. 4. Architecture overview of the developed coverage and capacity planning framework including network slicing planning and overprovisioning.

II. Related Work

Related Work regarding automated network planning and configuration is detailed in [3] and [5]. However, further work has been done in this field of research since then. In [8] and [9], the authors utilized unsupervised ML and clustering-based network planning, further confirming the viability of this method for this purpose. However, our work extends these ideas by allowing the usage of complex ray-tracing simulations instead of pure analytically calculated data basis and the integration of automated network slicing-based capacity planning. Regarding automated capacity planning, the authors in [10] present a suiting method for multi-tenant small cells. In our work, we aim to further extend these approaches with the novel aspects of network slicing as well as uRLLC and data predictability.

III. Methods

In this section, the overall architecture of the coverage and capacity planning framework will be described on the basis of Fig. 4. The figure is divided into four sections reflecting the chronological sequence and phases of network planning, from left to right *User Inputs*, *Target Signal Quality Calculation*, *Data Rate Calculations*, and finally *Coverage Planning*.

In the first section (*User Inputs*) on the left the parameters are listed, which have to the entered by the user of the framework. Firstly, the polygon of the private network is defined and serves as input of the coverage planning framework (cf. [3]), which downloads the necessary environment data from open sources and generates the Radio Environmental Maps (REMs) and performs the coverage planning based on it. However, in order to perform this coverage planning, a target value of the required received power (in dBm) is necessary, which was defined manually until now. This work aims to calculate the required receive power based on the network configuration as well as the target data rate (in Mbps). The network configuration consists of all different parameters, which directly influence the resulting data rate or throughput of the UEs, e.g., bandwidth, frequency, Time Division Duplex (TDD) pattern and many more (cf. Fig. 4 and Table I). As for the target data rate in Mbps, the slicing configuration is considered. This is defined as follows:

$$\textit{Target Data Rate (Mbps)} = DR_{eMBB} + (DR_{uRLLC} * (1 + \gamma_{OP})) \quad (1)$$

where γ_{OP} describes the overprovisioning factor, and DR_{uRLLC} and DR_{eMBB} define the peak resource allocation at any given time by the uRLLC and eMBB slice, respectively.

The next section (*Target Signal Quality Calculation*) describes the process of calculating a target signal quality by estimating possible Signal-to-Interference-plus-Noise-Ratio (SINR) values. The SINR calculation is conducted as follows:

$$SINR_{\mathrm{dB}} = SNR_{\mathrm{dB}} = P_{S,\mathrm{dBm}} - P_{N,\mathrm{dBm}} \quad (2)$$

where P_S is the received signal power, and P_N is the noise level, both with the unit dBm.

The following assumptions are made:

- As this work is in the domain of private 5G networks, the interfering signal level P_I is neglected, because neighboring private 5G networks are rare. Additionally, the regulations make sure that interference is kept low.
- All gains are assumed to be $0\,\mathrm{dB}$, as all powers are declared as Equivalent Isotropically Radiated Power (EIRP).

The noise level P_N is represented by the thermal noise P_T at the receiver, which is calculated as follows:

$$P_{N,\mathrm{dBm}} = P_{T,\mathrm{dBm}} = -174 + 10 * \log_{10}(B) \quad (3)$$

where B is the bandwidth in MHz.

As for P_S, all possible received power levels are considered in order to find the matching value for the target data rate. To do this, resulting data rates must be calculated from the SINRs.

This is conducted in the next phase *Data Rate Calculations*. For this, a corresponding Channel Quality Indicator (CQI) has to be derived from the SINR, for which no direct method exists in the 5G standard [7]. Consequently, existing work in the area of SINR to CQI conversion is referenced. In [11], the authors derive such tables for *64QAM* and *256QAM* modulation in the downlink. The goal is to use the best CQI possible, while maintaining a maximum transport block error probability or Block Error Rate (BLER) of 0.1 (or 0.0001 for low spectral efficiency tables). Derived from this, these tables are utilized based on the following assumptions:

- The same CQI used in the downlink is also used for the uplink, as CQI tables 1-3 in the 5G standard [12] are shared between Physical Uplink Shared Channel (PUSCH) and Physical Downlink Shared Channel (PDSCH) [7].
- For both downlink and uplink, the *256QAM* table is used, which enables the maximum throughput possible.
- The "Practical channel estimation" is preferred over "Perfect channel estimation", as it better represents real conditions.

After all possible SINR values are mapped to a corresponding CQI value, a list of possible code rates and modulation orders can be derived. This means that all required parameters are now present in order to calculate the resulting throughput for each received signal strength. The throughput or the max. data rate formula (same for both uplink and downlink) is given in the 5G standard as follows [7] [13]:

$$\textit{Data rate (Mbps)} = 10^{-6} * \sum_{j=1}^{J} \Big[v_{layers}^{(j)} * Q_m^{(j)} * f^{(j)} * R_{max} * \frac{12 * N_{PRB}^{BW(j),\mu}}{T_s^{\mu}} * (1 - OH^{(j)}) \Big] \quad (4)$$

A detailed description of the expressions can be found in [7] and [13].

For TDD, Eq. 4 has to be multiplied by a factor $\tau_{TDD}^{(Direction)}$ for each uplink or downlink direction to include TDD pattern:

$$\tau_{TDD}^{(Direction)} = \frac{1}{14} * N_{TDD}^{(Direction)} \quad (5)$$

where $N_{TDD}^{(Direction)}$ is the number of slots used for either direction (F slots included).

The throughput calculations result in received power level and throughput tuples. Based on the target data rate, a received power in dBm can be derived, which fulfills the requirement. On this basis, the coverage planning described in [3] is then performed. After that, the resulting data rates of the REMs are then calculated based on the same methods described above, as the utilized ray-tracing tool outputs received power levels.

In the next section, this method is further evaluated based on scenarios combined from [3] and [6].

TABLE I
CONFIGURATION PARAMETERS FOR THE NETWORK PLANNING, RAY TRACING, AND NETWORK SLICING DOMAINS

General Network Configuration	
Communication Direction	Uplink
Center Frequency	$F = 3.75\,\text{GHz}$ (FR1)
Bandwidth	$BW = [20, 50, 80, 100]$ MHz
TDD Uplink Slots	$N_{TDD}^{(Uplink)} = [4, 5, 6, 7, 8, 9, 10]$
Subcarrier Spacing	$\mu = 1$ (30 kHz)
Aggregated Carriers	$J = 1$
MIMO Layers	$v_{layers} = 2$
Communication Overhead	$FR1_{Uplink} = 0.08$ [13]
Scaling Factor	$f = 1$ [13]
CQI Table	Table 2 (256QAM) [7, Tab. 294]
SNR-CQI Table	256QAM [11, Tab. 2]

Network Planning and Ray Tracing Configuration (cf. [3])	
Overprovisioning Factors	$\gamma_{OP} = [0, 0.2, ..., 1, 1.2, ..., 2]$
Antenna EIRP	$P_A = [15, 18, ..., 39, 42]$ dBm
Antenna Radiation Patterns	Omnidirectional, Sector (120°)
Simulation Model	Standard Ray Tracing Model
Prediction Height	1.5 m
Max. Allowed Edge Power	−80 dBm

Network Slicing Configuration (cf. [6])		
Slice	**User Amount**	**Peak Resource Allocation [6]**
uRLLC	50	$DR_{uRLLC} = (1 + \gamma_{OP}) * 100\,\text{Mbps}$
eMBB	40	$DR_{eMBB} = 400\,\text{Mbps}$

IV. EVALUATION

In this section, the evaluation of the developed method of our network slicing-based coverage planning is presented based on a realistic scenario, which combines the properties of the scenarios in [3] and [6].

A. Evaluation Scenario and Configuration Parameters

In Fig. 5, the evaluation scenario is illustrated as a map generated by the network planning framework.

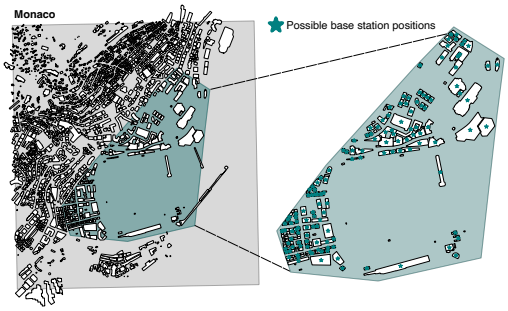

Fig. 5. Evaluation scenario based on a densely built-up area in Monaco. The stars represent possible base station locations.

There, a realistic evaluation scenario is shown, which was also utilized in [3] and is represented by a densely built-up area in Monaco. The teal area represents the private 5G network to be covered at the target data rate, where the received power must not exceed a certain value outside the area. All possible base station locations are represented by the teal stars,

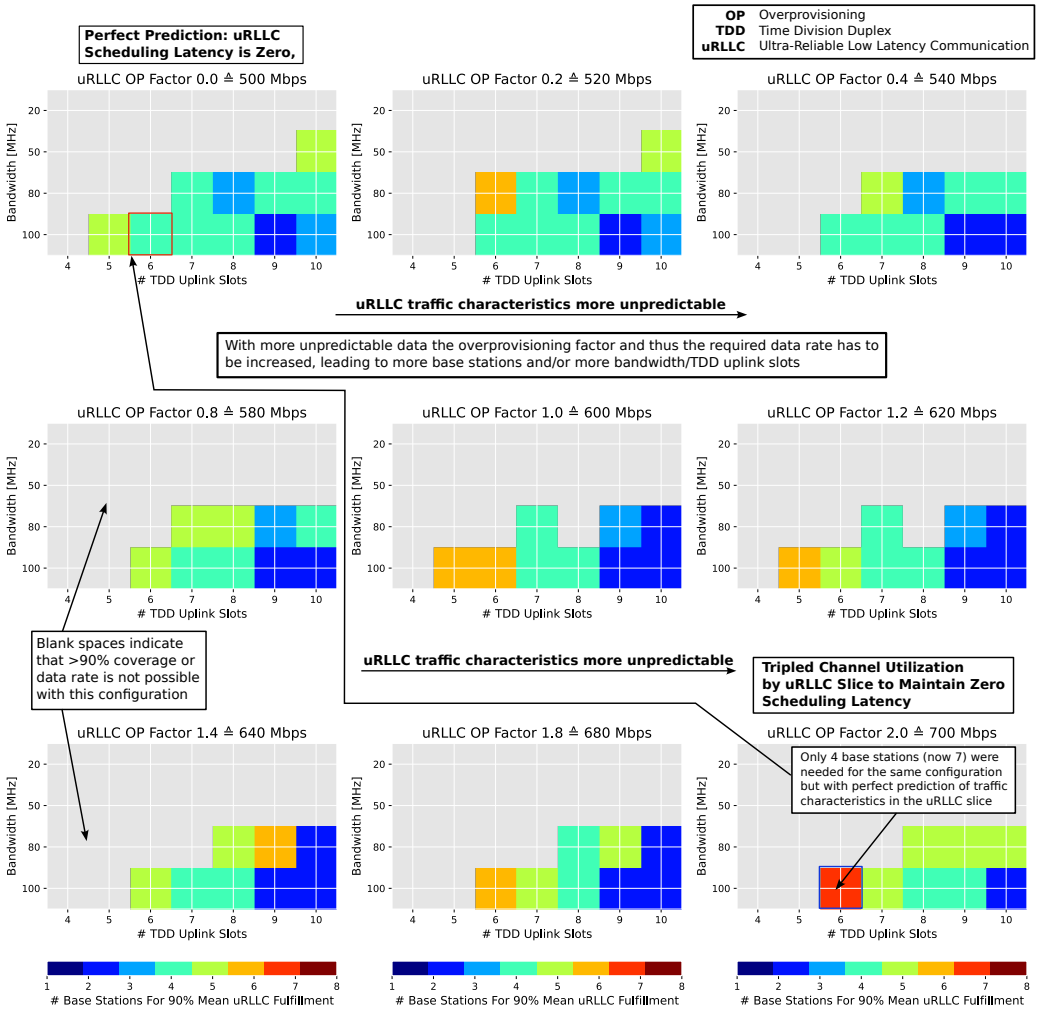

Fig. 6. Evaluation results of the coverage and capacity planning framework. Flawed prediction and thus high overprovisioning leads up to 3 additional base stations to be installed to keep zero latency in uRLLC slice, while strongly impairing the eMBB throughput due to overprovisioning.

which are mainly composed of positions on top of buildings with a minimal height and area. All configuration parameters for the procedure described in Sec. III are listed in Tab. I. Three different parameters are varied in order to analyze the influence of data predictability and thus overprovisioning on the network planning: bandwidth, TDD uplink slots, and most importantly the overprovisioning factor. The latter represents the parameter that reflects the predictability of the data, since poor predictability requires higher overprovisioning of the uRLLC slice. Planning is carried out for peak traffic (worst case scenario), whereby the network subscribers can be distributed anywhere on the area, i.e., also at the cell edge.

The results are shown and analyzed in the next section.

B. Evaluation Results and Analysis

In Fig. 6, all evaluation results are shown. The figure shows 9 different subplots for each Overprovisioning (OP) factor, starting on the top left with $\gamma_{OP} = 0$ (perfect prediction, no OP) and ending on the bottom right with $\gamma_{OP} = 2$ (strongly flawed prediction, tripled channel utilization). The subplots show the number of configured TDD uplink slots on the x axis, while depicting the chosen bandwidth (MHz) on the y axis. The color of the blocks plotted on the subplot each represent the number of base stations (or gNodeB) needed for $90\,\%$ mean uRLLC fulfillment. Fulfillment here means that the required number of resources (including OP) is available at the respective location. This is then provided in $90\,\%$ of the teal area shown in Fig. 5. The blank spaces indicate that $> 90\,\%$ coverage is not possible with the given configuration.

On the top left, where perfect prediction is assumed and no OP has to be configured, the required number of base stations is relatively low with 1 to 5 depending on the bandwidth and number of TDD uplink slots. The trend is, as expected, that the total of required base stations decreases with the increase of the bandwidth as well as the uplink slots. Note that ML-based automated network planning was performed and due to statistical fluctuations, the required base stations can increase slightly even when expanding bandwidth or TDD uplink slots.

The subplots further to the right show cases in which the uRLLC slice traffic characteristics become less and less predictable and thus the OP factor γ_{OP} must be increased. It can be observed that with decreasing predictability, the

average number of base stations required for the respective configurations also increases. The subplots further show which trade-offs between bandwidth, TDD configuration and slice predictability can be weighed against each other and selected by network operators.

This trend resumes in the second and third rows (beginning with the lower γ_{OP} on the left). The most interesting insight is revealed when analyzing the last subplot at the bottom right, which shows case $\gamma_{OP} = 2$. It shows the situation in that the traffic predictability is so low that the triple channel utilization is configured for the uRLLC slice and thus there is almost a static provisioning of RBs. In particular, the configuration 6 uplink slots and $100\,\mathrm{MHz}$ bandwidth is interesting: if only 4 base stations were needed with perfect predicability, now 7 base stations are needed with this OP factor to provide the same quality of service for the uRLLC slice (i.e., maintaining zero scheduling latency). This is compounded by the fact that this severely compromises the throughput of the eMBB slice, as unused uRLLC resources are wasted (cf. Sec. I-B).

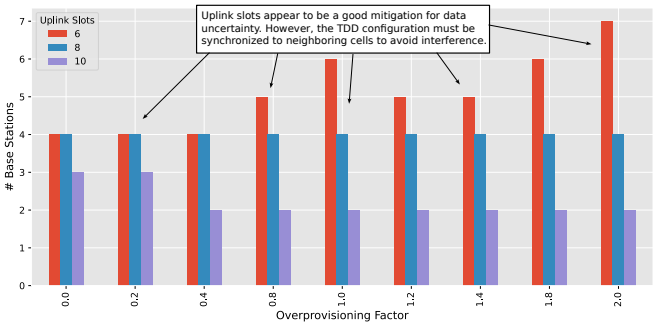

Fig. 7. Number of required base stations over the OP factor for $100\,\mathrm{MHz}$ comparing 6, 8, and 10 TDD uplink slots.

In Fig 7, a detailed comparison of $100\,\mathrm{MHz}$ bandwidth as well as 6, 8, and 10 uplink slots are given for all OP factors. It can be seen that a considerable trade-off exists between the number of TDD uplink slots and the number of required base stations. Moreover, for 6 uplink slots it becomes especially visible that overprovisioning (or data predictability) has a clear impact on the number of base stations required. However, the number of TDD slots often cannot be freely chosen, since they have to be aligned with the neighboring cells. Accordingly, increasing the number of base stations is the only free parameter that can be influenced by the network operator, but this is associated with higher costs.

V. Conclusion and Outlook

In this work, we presented a automated framework for network slicing-based network planning, which extended our previous work regarding rapid ML-based coverage planning. The capacity planning framework includes all configuration parameters given by the 5G standard as well as additional factors like TDD slot configurations. We showed that not only different parameters like bandwidth and slot configurations influence the network planning results and interesting trade-offs can be considered, but also the data uncertainty of uRLLC slices. Overprovisioning is required, when data can not be predicted well, e.g., when data is too random or scarce. This, in effect, wastes resources and limits the spectral efficiency and the throughput of other slices, e.g., eMBB applications. So, in summary, network operators need to consider data predictability in network slicing-based capacity planning, and this affects the physical layout and number of base stations, and thus also greatly increases the cost to meet requirements.

As for future work, more scenarios and influences of other network configuration parameters like Subcarrier Spacing (SCS) and carrier aggregation can be considered. Additionally, the system can be validated using real-life measurements or simulation frameworks like *ns-3*.

Acknowledgment

This work has been supported by the Ministry of Economic Affairs, Industry, Climate Action and Energy of the State of North Rhine-Westphalia (MWIKE NRW) along with the projects *Plan&Play* under the funding reference 005-2008-0047, the *5Guarantee* project under grant number 005-2008-0077, and the *Competence Center 5G.NRW* under grant number 005-01903-0047. Additionally, this work was supported by the German Federal Ministry of Education and Research (BMBF) in the course of the 6GEM research hub under grant number 16KISK038.

References

[1] A. Aijaz, "Private 5g: The future of industrial wireless," *IEEE Industrial Electronics Magazine*, vol. 14, no. 4, pp. 136–145, 2020.

[2] J. Rischke, P. Sossalla, S. Itting, F. H. P. Fitzek, and M. Reisslein, "5g campus networks: A first measurement study," *IEEE Access*, vol. 9, pp. 121 786–121 803, 2021.

[3] C. Bektas, S. Böcker, B. Sliwa, and C. Wietfeld, "Rapid network planning of temporary private 5g networks with unsupervised machine learning," in *2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall)*, 2021, pp. 01–06.

[4] N. H. Mahmood, S. Böcker, I. Moerman, O. A. López, A. Munari *et al.*, "Machine type communications: key drivers and enablers towards the 6g era," *EURASIP Journal on Wireless Communications and Networking*, vol. 2021, no. 1, p. 134, 2021.

[5] C. Bektas, C. Schüler, R. Falkenberg, P. Gorczak, S. Böcker *et al.*, "On the benefits of demand-based planning and configuration of private 5g networks," in *2021 IEEE Vehicular Networking Conference (VNC)*, 2021, pp. 158–161.

[6] C. Bektas, D. Overbeck, and C. Wietfeld, "SAMUS: Slice-aware machine learning-based ultra-reliable scheduling," in *ICC 2021 - IEEE International Conference on Communications*, 2021, pp. 1–6.

[7] C. Johnson, *5G New Radio in Bullets*. Independently Published, 2019.

[8] M. Chraiti, A. Ghrayeb, C. Assi, N. Bouguila, and R. A. Valenzuela, "A framework for unsupervised planning of cellular networks using statistical machine learning," *IEEE Transactions on Communications*, vol. 68, no. 5, pp. 3213–3228, 2020.

[9] M. Umar Khan, M. Azizi, A. García-Armada, and J. J. Escudero-Garzás, "Unsupervised clustering for 5g network planning assisted by real data," *IEEE Access*, vol. 10, pp. 39 269–39 281, 2022.

[10] P. Muñoz, O. Sallent, and J. Pérez-Romero, "Self-dimensioning and planning of small cell capacity in multitenant 5g networks," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 5, pp. 4552–4564, 2018.

[11] A. K. Thyagarajan, P. Balasubramanian, V. D, and K. M, "Snr-cqi mapping for 5g downlink network," in *2021 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob)*, 2021, pp. 173–177.

[12] 3GPP, "Physical layer procedures for data," 3rd Generation Partnership Project (3GPP), TS 38.214, Release 16.8.0, Tech. Rep., Dec. 2021.

[13] ——, "User Equipment (UE) radio access capabilities," 3rd Generation Partnership Project (3GPP), TS 38.306, Release 16.8.0, Tech. Rep., Mar. 2022.