Capacity Optimization for Ultra-Reliable Low-Latency Communication in 5G - The SON Perspective

Elke Roth-Mandutz*, Abubaker-Matovu Waswa†, Andreas Mitschele-Thiel†
*Fraunhofer IIS, Erlangen, Germany
†Integrated Communication Systems Group, Technische Universität Ilmenau, Ilmenau, Germany
Email: elke.roth-mandutz@iis.fraunhofer.de, [abubaker-matovu.waswa, mitsch]@tu-ilmenau.de

***Abstract*—Towards 5G, the challenge to achieve high reliability and low latency asks for new directions in systems so far strongly focusing on high data rates. Schemes to achieve high reliability for low latency demanding services are link diversity to ensure reliability and resource reservation to avoid scheduling delays. However, these schemes place high demands on the scarce radio resources. In this paper we propose a self-optimized approach on the network management level to minimize the capacity impact for ultra-reliable low latency communication (URLLC) services. The idea of the new URLLC Self-Organized Network (SON) is to provide optimized sets of parameters derived from the respective service requirements and the current conditions in the network. We discuss Network Management (NM) parameters regarding their impact on URLLC and present a structure, which allows a fast selection of the appropriate parameter set. One focus is on device-to-device (D2D) communication, which has a high potential to meet both, the reliability and latency requirements. An initial study on radio resource reuse indicates a promising gain in network capacity, while D2D reuses the same resources as the cellular users in the network.**

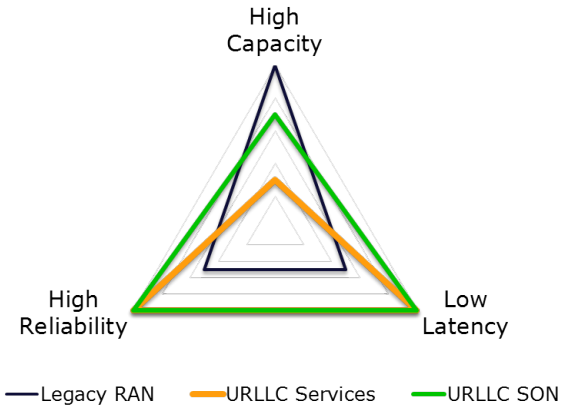


Figure 1. Latency - Reliability - Capacity Dependency: Capacity Impact on Legacy RAN at the Expense of Latency and Reliability for URLLC Services without and with SON Support

I. Introduction

High reliability and low latency demanding applications are some of the major challenges for the 5th generation cellular networks. These applications fall into the service category ultra-reliable and low-latency communication (URLLC), which was defined during the recent 5G standardization work [1]. Important examples for URLLC are automated traffic control in the automotive area and mission critical machine type communication in factories. Reaching the targeted end-to-end latency of at most 1 ms as well as the reliability of 10^{-9} (maximum probability of packet delivery failure) poses challenges, which are mostly investigated on the physical layer (e.g. [2]), but not yet considered at the network management level.

Ultra-reliable low latency applications assume highest priority to reach the requested latency for a given reliability - even at the cost of capacity. An impact on capacity is expected due to the relations between capacity on the one hand and latency and reliability on the other hand [3]. One approach to meet the strict latency requirements is the reservation of resources to avoid any scheduling delay. Additionally, link diversity, i.e. concurrent transmission of the same data over multiple radio links, is proposed to fulfill the reliability target. As a result URLLC poses extra cost on capacity, which contradicts the common capacity maximization objectives of legacy radio access networks (RAN).

Figure 1 demonstrates the relationship and dependency of latency, reliability and capacity in the RAN. The blue triangle illustrates the clear preference towards capacity in legacy RAN networks, giving a low priority to latency and reliability. With the implementation of URLLC, the priority shifts towards low latency and high reliability at the expense of capacity. The expected major drawback on the capacity is indicated by the orange triangle in figure 1. Implementing the new URLLC Self-Organized Network (SON), a better balance between the 3 competing objectives is achieved: while guaranteeing the latency and reliability requirements, the capacity for any given set of conditions is optimized (see green triangle in figure 1). The optimization considers the particular environmental conditions including geographical, load, and mobility factors.

5G services demanding ultra-reliability (UR) and low latency are illustrated in figure 2 highlighting the relation between reliability and latency for selected services [4], [5]:

- **Automotive**: Vehicle-to-everything (V2X) e.g. automated traffic control including intersection management and fleet driving to mitigate road accidents and improve traffic efficiency towards fully autonomous driving

978-3-901882-98-2

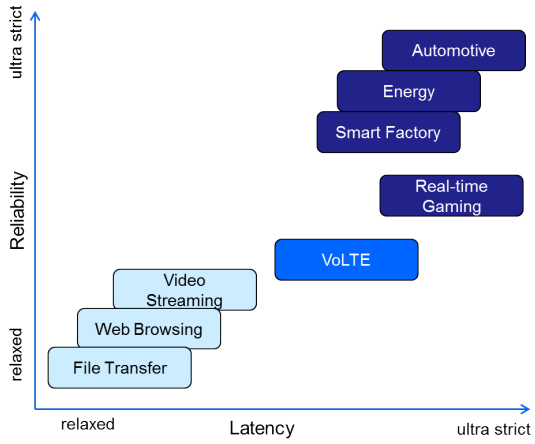


Figure 2. Example Services with Varying Latency or Reliability Requirements

- **Smart factory**: Massive Machine-Type Communications (mMTC) for collaborative robots to control industrial applications
- **Energy:** Automatic energy distribution, monitoring and control of smart grids
- **Real-time gaming**: Interactive on-line games

The consequences of not fulfilling the latency or reliability requirements depend on the particular application. For example, in case of video streaming latency and reliability constraints are almost negligible, while too high latency results in a major quality impact for real-time gamers (see figure 2). Meeting the demanded latency and reliability is however crucial for safety-critical applications.

In literature, several approaches to minimize the latency while ensuring high reliability in future 5G networks are discussed, [6], [7], [8]. 3GPP investigates the network latency and diverse approaches for latency reduction in [2]. Few new approaches to optimize radio resource management (RRM) for latency reduction were proposed. One new RRM concept called, Transmission Time Interval (TTI) shortening, allocates resources below standard sub-frame size basis. This concept shortens the minimum length of the legacy TTI from 1 ms to 0.5 ms or less on an orthogonal frequency division multiplexing (OFDM) symbol basis [2], [10]. In addition, the concept of instant or fast uplink access avoids any scheduling delay by reserving resources for URLLC use cases [10]. Reliability improvements e.g. by link diversity are discussed in [6] and [11]. Finally [3] investigates the relation and impact of URLLC on capacity.

Device-to-device (D2D) communication is one promising approach to significantly reach the ambitious goal of 1 ms [12] to virtually zero latency [13]. User equipments (UEs) in the vicinity are enabled to communicate directly with each other without involving the evolved NodeB (eNB). In addition to the proximity and hop gain in D2D, resource management can be optimized not only to reduce the latency, but also to increase the system capacity. The capacity gain can be achieved by ab efficient radio resource allocation and sharing strategies to integrate D2D within the existing cellular network. The approach of spatial reuse of radio resources between D2D pairs with cellular UEs (C-UEs) or other D2D-pairs within the same cell is further investigated in section V.

At the same time, the variety of services managed simultaneously in 5G networks increases the network heterogeneity and management complexity, asking for a new level of self-organized solutions. The new URLLC SON use case is proposed as one important contribution to ensure the ambitious latency and reliability claims as new objectives for 5G network optimization.

In this paper, we concentrate on an initial step to find out, which of the NM parameter settings meet the URLLC requirements at minimal cost, i.e. which combination of techniques to apply (e.g. robust modulation and coding, higher transmission (Tx) power, link diversity). Thus, we aim at provisioning an optimized set of network management parameters for any particular environmental network condition to increase the capacity while serving the latency and reliability requirements. With the URLLC SON use case we expect conflicts with other use cases, which need to be resolved.

The specific **contributions** of this paper are:

- Propose a self-organized use case for URLLC with focus on D2D at the network management level.
- Provide an optimal set of system parameters for different environmental conditions, i.e. maximize the capacity while meeting latency and reliability constraints.
- Propose a hierarchical organization of the optimized parameter sets to reduce the potentially huge number of sets and accelerate the access to the sets.
- Demonstrate an approach on D2D interference indicating the potential of spatially distributed intracell radio resource reuse.

The remainder of this paper is organized as follows: Section II defines a set of SON requirements for network slices and services in 5G networks demanding URLLC. The SON architecture and processing are then defined in Section III. Next, in Section IV we propose approaches for a self-optimized URLLC use case for D2D, including the identification of network management parameters in relation to the particular environmental conditions. In section V, we present D2D interference scenarios to demonstrate the effect of spatial radio resource reuse. Finally, we conclude the paper with an outlook in Section VI.

II. Service Concept and Parameters

URLLC demands priority on reliability and low latency over capacity to reach the ambitious URLLC goals, which results in conflicts between capacity maximization as pursued by the legacy networks. One concept to address the conflicting requirements is network slicing. In addition, as the cost in terms of capacity for URLLC is expected to be very high, a careful definition of latency and reliability is needed.

A. Network Slices

Within 5G, the recently introduced network slicing provides a more flexible architecture, to separate the wide range of applications with different needs. Several network slice instances may be used concurrently by one RAN, where each slice has its own set of policies and NM parameters [1]. Network slicing allows to support several different logical networks on the same physical network infrastructure. Thus, a reduction in cost and energy consumption to run a network compared to deploying separate physical networks for the different use cases or business scenarios is expected [14]. Each slice maps completely - or at least to a large extent - to one of the 3 service categories [15] for which future 3GPP new radio (NR) solutions are currently specified:

1) Enhanced Mobile Broadband (eMBB): Demanding higher data rates for applications such as video streaming, web browsing, and file transfer,
2) mMTC: Extending the LTE Internet of Things capabilities to support huge numbers of mostly wearable, low cost devices with enhanced coverage, and long battery life,
3) URLLC.

B. Ultra-high Reliability

Reliability is defined as the capability of guaranteeing a successful message transmission to a defined high degree. An impact on reliability is experienced, if either a message is lost, a message is too late or a message experiences residual errors.

Solutions to achieve the required ultra-high reliability include diversity, e.g. simultaneous transmission on multiple links on different frequencies or radio access technologies (RATs), as well as antenna diversity.

C. Low Latency

In the scope of this paper, latency is defined as the time between data being generated from one device and the same data being correctly decoded by another device. For example, 3GPP TR 38.913 targets for the user plane a latency of 0.5 ms for downlink as well as for uplink.

To reach the strict low latency, new concepts on the physical and MAC layers are introduced. It is the admission control's responsibility to provide the priority according to the given quality of service (QoS) or other priority settings. In [2] and [7], the key features to achieve the latency goals are seen as a reduced processing time and shorter periodicity for resource transmission. Approaches to reduce the latency on different layers include:

- Avoid scheduling delay: Reserving resources, also referred to as "semi-persistent scheduling" or "fast uplink access" is required to ensure immediate resources allocation.
- Short periodicity and fast processing for small packet size: Reduced TTIs
- No or very fast retransmission on the MAC layer: Depending on the latency requirements, Hybrid Automatic Repeat Request (HARQ) needs to be strongly accelerated or fully omitted.
- Preference of direct D2D communication over device-to-infrastructure (D2I), assuming a major proximity and hop gain.
- Preference for URLLC: Current 3GPP networks support QoS mechanisms. With the QoS class indicators (QCI) different service applications can be prioritized using allocation and resolution priority values [14].

III. URLLC SON Architecture and Assumptions

In this section we introduce the architecture of the URLLC SON, followed by a set of assumptions for our proposed SON approach.

A. Architecture for the URLLC SON Use Case

In general, SON is categorized in 3 different types of architectures: First, the centralized architecture, implementing the SON algorithms on a higher level network management system, second the distributed architecture with SON algorithms at the eNB and third, the combination of both, the hybrid architecture. For the URLLC SON, we propose a hybrid architecture to take advantage of the distributed approach, allowing a faster reaction and reduced data exchange and the centralized approach where the interaction between multiple eNBs enables advanced learning mechanisms. Figure 3 presents our proposed hybrid URLLC architecture including the data flow between the nodes. The involved network nodes are the NM system, which implements the SON algorithm, the eNB together with its dedicated UEs, and the Core Network (CN). Input for SON is retrieved from the NM, which provides the operator preferences (see 1 in figure 3), e.g. to allow D2D or limit the maximum number of links for the same connection. Next (2 in figure 3), the UEs and eNB provide the measurements, which are preprocessed by the eNB. Last, the network slice and/or service specific parameters are retrieved from the CN (see 3 in figure 3).

Then, using UE and eNB measurements as well as the CN service specific requirements as input, the eNB derives the environmental profile. The current environmental profile reflects the radio conditions by summarizing the currently experienced geographical, mobility and traffic situation. In addition, the eNB determines the current Key Performance Indicators (KPI) to be sent to the NMS. In case of parameter changes, the eNB transmits to the NM an updated set of the environmental profile and KPIs (see 4 in figure 3). On reception, the URLLC SON algorithm in the NM selects the appropriate optimized set of NM parameters to maximize the capacity for any given latency and reliability constraint.

B. Assumptions

URLLC applies to both types of connections, D2D and D2I, where D2D is given preference, if applicable. The reason is the expected better overall latency for D2D, caused by the direct communication between the UEs. Multiple hops between the network nodes, including delays due to data processing per

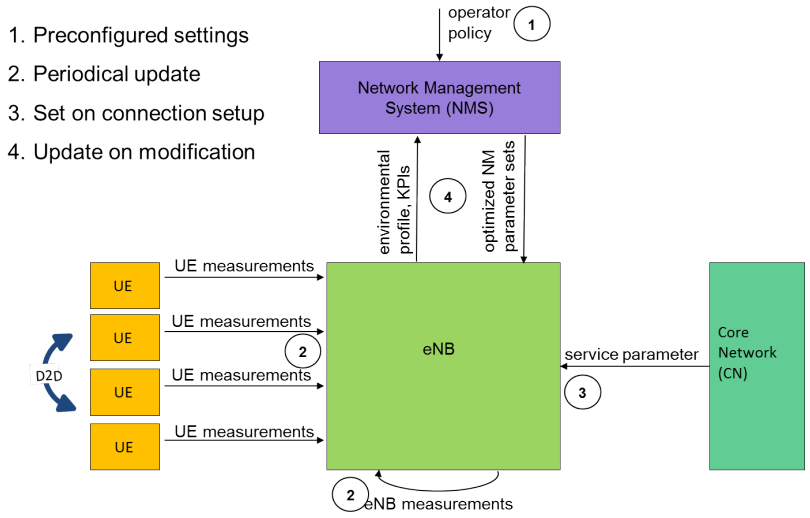


Figure 3. Hybrid SON Architecture and Message Flow for URLLC

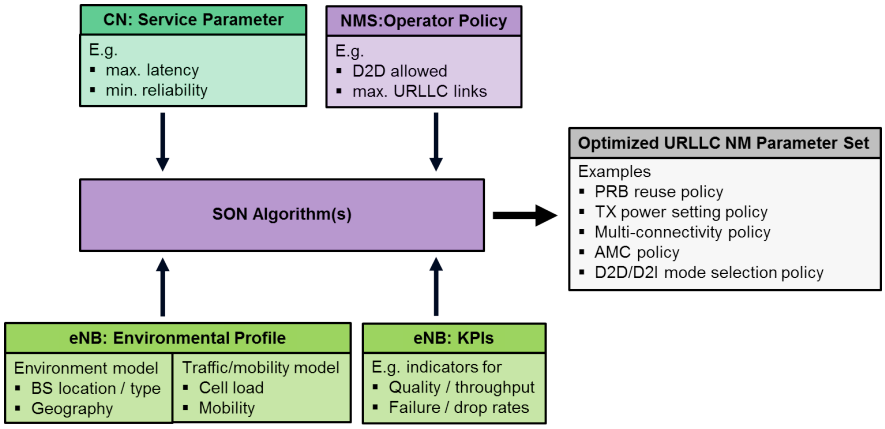


Figure 4. Set of URLLC NM Parameters to Optimize Capacity using Service Parameters (from CN), Operator Policy and KPIs (from NM) and the Environmental Profile (from eNodeB) as Input

node, is avoided when D2D communication is used. Additionally, we consider the following set of assumptions for URLLC services:

- Typically URLLC requires very small sized data packets, i.e. small payload.
- A bulky message flow is expected, e.g. high traffic within short time period due to critical traffic situation in V2X.
- Fulfilling the strict latency and reliability requirements demands new approaches on the lower layers as listed in section II.
- Very low latency does not allow any retransmission, i.e. HARQ is omitted.
- Only the licensed spectrum is considered.

IV. URLLC SON Approach

In this section we focus on 2 major aspects of the URLLC SON use case: First, on the URLLC SON parameters, and second, on the approach to select the most appropriate parameter set for given environmental conditions.

A. URLLC SON Parameters

The URLLC SON use case tunes NM parameters to optimize the resource utilization for services with given reliability and latency requirements. NM parameter settings apply on medium to long term basis, i.e. minute to hourly basis, different to the near real time adaptations required in the RAN. The task of NM parameters is to align RAN parameters towards higher level requests, e.g. service and network requirements and operator policies. Depending on the radio and traffic conditions and considering the given service requirements, the NM parameter settings are then optimized. Table I presents a list of NM parameters with impact on latency (LL), UR or both. In case an impact is existent, it is indicated with "high", otherwise "low" indicates no significant impact. The listed parameters are seen as an initial set, which will increase with further specifications in 5G.

1) PRB Reuse Policy:

Radio resource sharing policy controls the reuse of the

Table I
URLLC SON PARAMETERS: IMPACT ON LATENCY / RELIABILITY

PARAMETER NAME	LL	UR
PRB Reuse Policy	low	high
UE Tx Power Setting Policy	low	high
Multi Connectivity Policy	low	high
AMC Policy	high	high
D2D Mode Selection Policy	high	high

physical resource block (PRB), i.e. the PRB reuse factor. It defines, how often one PRB is allowed to be reused for D2D within the same cell. The range of the PRB reuse factor starts from "0" (i.e no reuse - overlay D2D) and is typically limited by the interference caused by intra-cell PRB reuse (underlay D2D). Increasing the reuse factor is expected to increase the interference risk, resulting in an impact on the reliability as indicted in table I. Reuse of resources may be based on multiple schemes, e.g. the distance of the D2D-UEs (i.e. spatial reuse) or by thresholds using the measured Signal-to-interference-plus-noise ratio (SINR) to restrict the resource reuse. We provide one scenario of resource reuse for D2D in section V. In addition, the UE mobility impacts the reuse factor: Resource reuse for medium to high speed UEs is assumed to cause more interference. The reason is that the SINR to UEs in vicinity changes rapidly. Moreover, the restricted PRB reuse areas in a cell force a fast and continuous assignment of new PRB. Thus, high speed UEs demanding UR may be restricted from resource reuse.

2) Tx Power Setting Policy:

Tx power increase is expected to result in a higher SINR and consequently, in an improved reliability of the UE's radio link (see table I). On the other hand the Tx power and corresponding interference increase may negatively impact the transmission quality of any other UE in vicinity. Typical legacy networks aim firstly on maximizing the resource utilization. One contribution is to minimize interference by low Tx power usage, which in addition serves the secondary goal, to save UE battery power. As a consequence, the Tx power policy objective is to assign the minimum UE Tx power fulfilling the requested quality demand. For URLLC, this legacy policy does no longer apply. Giving priority to UR, an increase in Tx power is implicitly accepted, surpassing the legacy network objectives. The examples for URLLC Tx power policy list few initial rules for URLLC, which may result in assigning higher Tx power than required according to measurement results at the given point in time:

- A minimum Tx power threshold may restrict the UE from low power settings.
- During URLLC call setup phase, the UE Tx power control uses the maximum Tx power.
- An extended period for Tx power reduction in the RAN power control algorithm is assumed to avoid power reduction on a very short term basis.

The Tx power policy for URLLC in D2I applies also to D2D, which uses the uplink radio resources. However for D2D, power control is restricted in the current 3GPP specification [16].

3) Multi Connectivity / Diversity Policy:

The multi-connectivity or radio link diversity policy aims on improvement of the reliability (see table I) by establishment of multiple radio links for one connection. Multiple radio links can either be established within the same RAT (single RAT) or in different RATs. Several types of multi-connectivity are distinguished, [8], [17]:

- 5G-LTE Dual Connectivity: The 5G UE moves between LTE and 5G radio access coverage areas, establishing simultaneous connections with both networks before seamlessly handing over. 5G-LTE Dual Connectivity will enable 5G networks to provide multi-standard and multi-band support.
- 5G Multipoint Connectivity: The 5G UE connects to 2 5G base stations simultaneously, improving bit rate performance through multiple downlink streams, as well as the signal strength and resilience. 5G multipoint connectivity will be key to supporting multi-layer networks with macro and small cell coverage.
- MIMO: Massive multi-input multi-output (MIMO) antennas can provide reliable links by benefiting from spatial diversity. For example, beam-forming may mitigating effects of fast fading.

As NM parameter a threshold can determine the minimum number of radio links required for given environmental conditions. For example, in case the minimum radio link threshold for a given high traffic situation is set to "2", any new URLLC connection will immediately establish a minimum of 2 radio links.

For D2D connections, a secondary link can be established in infrastructure mode, expecting an increase in reliability. However, a latency increase is expected in D2I as detailed in section III-B.

4) Adaptive Modulation and Coding (AMC) Policy:

AMC policy refers to parameters limiting the modulation and coding schemes to those promising more robust transmission, but more delay (see table I). In general, more robust low order modulation improve the reliability, with negative impact on the throughput. For coding, HARQ is appropriate for eMBB, which ensures higher reliability, but also higher delay due to data re-transmissions on request. On the other side, any re-transmissions causes significant delay causing a serious risk for low latency requesting applications. For URLLC, modulation may therefore be restricted to low order modulation, e.g. QPSK or 16QAM,while HARQ may not be appropriate. Instead, the same data may be repeated either in sequence on the same radio link or quasi simultaneously on different radio links (see multi-connectivity above).

For D2D the same AMC approach applies, which is further limited and detailed by standardization constraints [16]. However, different to D2I, D2D may allow retransmission and still meet the stringent latency requirements. The main reason is

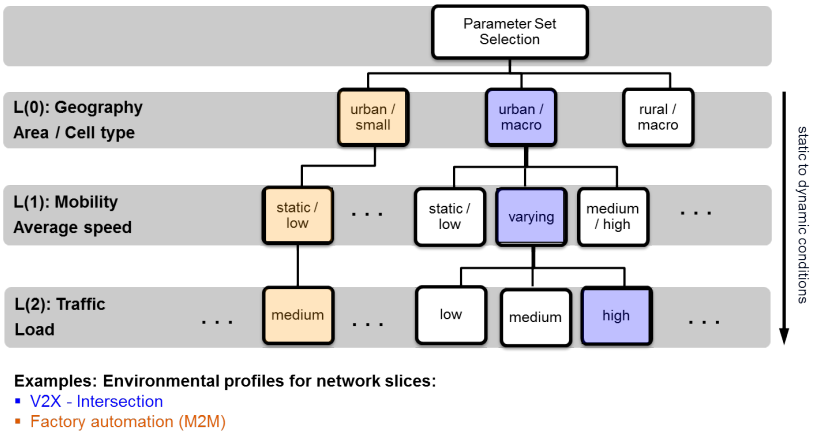


Figure 5. Hierarchical Structure of Environmental Profiles: From Static to Dynamic Conditions

the strongly reduced latency for direct D2D communication with no intermediate nodes compared to D2I.

5) D2D / D2I Mode Selection Policy:

If the communicating UEs are in proximity, direct D2D communication is more promising to meet the strict latency and even reliability requirements (see table I). The major reason for D2D preference is the reduced latency due to direct communication, avoiding multiple hops between the network nodes. Two further criteria impacting the mode selection are the velocity and the direction of motion of the UEs. For example, fast moving UEs in opposite directions may cause handover instantaneously, putting a risk on ensuring the required reliability and latency. Finally, the cell load impacts the mode selection as D2D enables the reuse of resources and less interference due to the reduced Tx power.

6) Coordination with legacy SON Use Case:

In addition to the network parameter policies, the URLLC SON use case is expected to interfere with several of the legacy SON use cases and vice versa. An update of most legacy SON use cases is required to allow URLLC services. We restrict the illustration of the impact of URLLC to legacy SON use cases to 2 examples as a more detailed analysis of the relation between the URLLC and all SON use cases is beyond the scope of this paper.

- The energy saving SON intends to deactivate cells, with very minor or no traffic over a given time period. UEs may therefore be forced to establish connections to active cells offering sub-optimal radio conditions. This behavior may jeopardize both, low latency and high reliability. Depending on the operator's priority, any cell or even neighbor cells with established URLLC connection, may therefore be excluded from the cell deactivation policy to meet the latency and reliability requirements.
- The aim of the Mobility Load Balancing (MLB) use case is to improve the system capacity by evenly distributing the load among cells. Depending on the cell load, handover parameters of the serving cell as well as the adjacent cells are adapted to move UEs from highly loaded cells to less loaded neighbor cells. However, for reliability reasons, URLLC need to remain in the cell with the strongest radio link. Thus, handovers for traffic offloading reasons need to be limited to non-URLLC connections, which may in turn demand to handover e.g. eMBB in an earlier phase to free off resources for additional URLLC traffic.

B. Hierarchical Structure for NM Parameter Sets

With the introduction of new services and network slicing in 5G, the complexity of NM functions to guarantee the optimal setting of the network parameters increases strongly. However, incorrect or sub-optimal configuration of network or UE parameters could strongly degrade the overall network performance. The assignment of the most appropriate parameters, meeting multiple environmental conditions at a given point in time for a given network slice requires the generation of a tremendous amount of different parameter sets; i.e. for each combination of environmental conditions and each network slice an optimized parameter set would need to be generated and maintained. Therefore, we propose a hierarchical structure, which models the frequency of changes for environmental conditions, to select the appropriate parameter set. We allow the reuse of the same parameter sets for multiple conditions. At the same time the proposed hierarchy accelerates and simplifies the identification of the best fitting parameter set.

The parameter set selection is initiated by the SON function, which monitors the KPIs (see subsection III-A). Whenever one of the KPIs reach a service specific threshold, a parameter set selection is executed, following the rules below:

1) Fulfill the Operator's policy
2) Ensure to use the environmental conditions at the given point in time
3) Select the most appropriate parameter set for the given environmental conditions

In figure 5, the hierarchical structure presents a top-down approach, which is grouped by the probability that a given condition may change. Each level represents one of the

following environmental conditions: Geography, mobility, and traffic. The upper level L(0) presents more static geographical conditions, including the area type (e.g. rural, suburban, urban) as well as the cell type (e.g. macro, micro, pico). Next, the second level L(1) provides the more flexible mobility conditions, including the average UE speed within the cell, varying from static or low speed to high speed. Depending on the time of the day and the cell location, the mobility conditions vary more or less frequently. For example, along a highway in rural areas, the mobility condition may remain rather stable, while cells in urban area may experience rapidly changing conditions depending on e.g. events, time of the day and location. Finally, the lowest level L(2) indicates the traffic load of the cell, which may vary strongly, e.g. depending on time of the day and day of the week.

We present 2 examples of network slices in figure 5: The top-down approach for the network slice V2X at an intersection is colored in purple. Starting the selection of the optimized parameter set from the top, we find a macro cell in an urban environment at L(0), with currently strongly varying average speed at L(1) at an intersection for high traffic load at L(2). As a second example, we present M2M communication used for factory automation colored in orange. Assuming an urban environment with small (indoor) cells, the typical M2M mobility is static to low moving (robots) with medium traffic load, while the factory is in operation.

The predefined parameter sets assume that the same optimized parameter values are valid for similar environmental conditions and can thus be reused for multiple scenarios.

V. D2D Resource Reuse Scenarios

The reuse of cellular radio resources for D2D communication (i.e. underlay D2D mode) is seen as one way to improve resource utilization and thus, enhances the network capacity. However, reuse of radio resources by both cellular UEs (C-UEs) and D2D UEs (D-UEs) leads to intra-cell interference which must be managed to keep it below given threshold levels. Additionally, safety critical services demand preferential treatment when reusing radio resources to guarantee their QoS (e.g. reliability).

One advantage of the required proximity of the D-UEs, is the significantly lower average Tx power compared to that of C-UEs. Thus, a significant level of interference caused by the D-UEs is assumed to be limited to a restricted area surrounding any given pair of D-UEs. Similarly for cellular communication, the area with high interference level depends on the Tx power of the C-UE and its location from the serving eNB. Reuse of the same radio resource is thus not permitted within this area, referred to as interference limited area (ILA) [18]. Spatial separation of the different ILAs to avoid their overlap ensures minimal intra-cell interference between UEs reusing the same radio resources. The main parameters that determine the extent of resource reuse within the cell are:

- Cell load
- SINR-threshold for the UEs

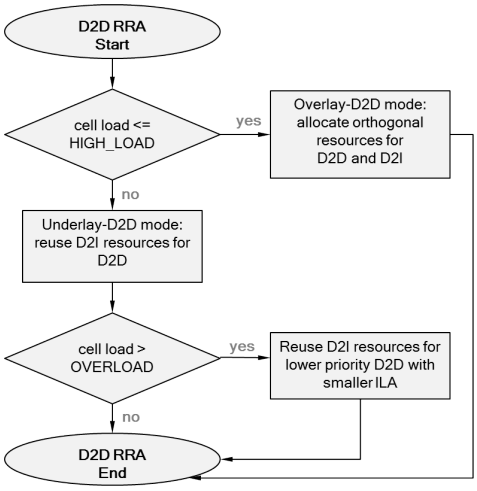
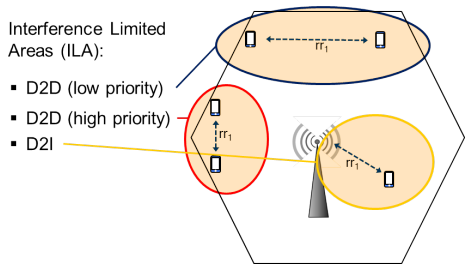


Figure 6. Cell Load based D2D Radio Resource Allocation (RRA)

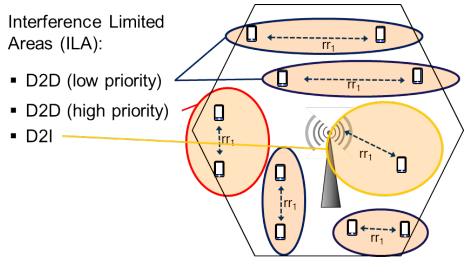
- Maximum Tx power for both C-UEs and D-UEs.

In figure 6 we present a flow chart for spatial radio resource reuse for D2D based on the cell load. Summarizing, we distinguish between 3 load conditions in a cell: low load, high load and overload with 2 thresholds to identify the boundaries between them. During low load conditions, adequate resources are available for all UEs (i.e. C-UEs and D-UEs) within the cell. Thus, no resource reuse is implemented i.e. orthogonal resources are used for both C-UEs and D-UEs (overlay D2D). In case the cell load hits the HIGH_LOAD threshold, there is a considerable risk that the cell runs out of resources. In this high load condition, the D-UEs should reuse the radio resources (i.e. underlay D2D), with sufficient spatial separation of their ILAs. Furthermore, in the case of overload, almost all resources within the cell are in use (possibly reserving a few resources for emergency purposes). The capacity shortage under overload conditions is mitigated by reusing resources for D-UEs having smaller ILAs. The size of the ILAs is proportional to the UEs' threshold SINRs and assigned maximum Tx power.

Figure 7 presents an example for radio resource reuse within the cell during high load and over-load conditions. In a given cell, it is expected that the D-UEs that could potentially share radio resources with a given C-UE will be randomly distributed throughout the cell. However, not all the D-UEs may be allowed to share the same resources with a given C-UE due to interference between them. In figure 7(a) 2 D2D pairs and one D2I connection are shown, where the circle indicates their respective ILAs, i.e. the area limited by a defined SINR threshold. Radio resource reuse is only allowed outside the ILA such that the SINR for the D-UEs is above the SINR-threshold. For this example, 2 priority classes for the D-UEs are considered: D2D high-priority for URLLC applications and D2D low-priority class for the other applications. The D2D high-priority class ILA is surrounded by a red, the D2D low-priority class ILA by a blue and the D2I ILA by a yellow frame. All D2I and D2D links use the



(a) High load conditions



(b) Over load conditions

Figure 7. Spatial radio resource reuse with the cell for both cellular and D2D communication

same radio resources (rr_1) and their ILAs should not overlap to avoid interfering with each others communication. In case of overload, demanding a higher resource utilization, 3 further D2D low-priority connections are allowed to reuse the same resource (rr_1) as presented in figure 7(b). To allow increased reuse of resources by D2D-pairs, the ILA size of the additional D2D pairs has to be smaller, i.e. higher interference is accepted or the maximum Tx power of the D-UEs is reduced. However, further resource reuse by D2D pairs with smaller ILAs is limited to D-UEs of the D2D low-priority class in contrast to the D2I and the D-UEs of D2D high-priority class, which remain unchanged.

VI. Conclusion

The introduction of URLLC is not only a major challenge on the RAN, but also on the network management. The proposed new SON use case is a high level optimization approach towards successful implementation of the strict low latency and reliability requesting services. We discussed NM parameters regarding their impact on URLLC services and the system capacity. In addition, we presented an approach to quickly select the most appropriate set of parameters. Adaptations to the legacy SON use cases are identified, which demand the distinction between the service types to meet the diverse requirements and ensure the best possible overall performance. As a preliminary study, we presented a scheme to optimize the reuse of radio resources for a mixed D2I and D2D traffic scheme.

As next steps, we intend to find out the conditions on which each NM parameter depends. Based on simulation studies we expect feedback from the system concerning conformance with URLLC requirements for given network conditions. Focusing initially on radio resource reuse, the parameter set will be optimized in a self-organized way, adding a learning capability for future enhancement.

Another step towards co-existence of URLLC with eMBB and mMTC applications and as well as D2I and D2D, is the adaptation of the legacy SON use cases defined in 3GPP. Until now, the legacy use cases focus mostly on optimization of capacity and mobility parameters for D2I. Service specific requirements need to be implemented by all use cases to further enhance parameter optimization in line with the more service- and user-centric 5G network.

Acknowledgment

This work has been supported by the German Federal Ministry of Education and Research (BMBF) under the project “fast-wireless”.

References

[1] NGMN Alliance, “NGMN 5G White Paper,” Feb 2015.
[2] 3rd Generation Partnership Project (3GPP), “TR 36.881: Study on latency reduction techniques for LTE (Release 14),” Jun. 2016.
[3] B. Soret, P. Mogensen, K. I. Pedersen, and M. C. Aguayo-Torres, “Fundamental tradeoffs among reliability, latency and throughput in cellular networks,” in *2014 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2014, pp. 1391–1396.
[4] METIS project, “Updated scenarios, requirements and KPIs for 5G mobile and wireless system with recommendations for future investigations,” *Deliverable D1.5*, April 2015.
[5] Ericsson, “5G Systems,” *White Paper*, Jan 2015.
[6] N. Johansson, Y.-P. Wang, E. Eriksson, and M. Hessler, “Radio Access for Ultra-Reliable and Low-Latency 5G Communications,” in *IEEE International Conference on Communication (ICC)*, Jun. 2015, pp. 1184–1189.
[7] O. N. C. Yilmaz, “Ultra-Reliable and Low-Latency 5G Communication,” Jun 2016.
[8] J. J. Nielsen and P. Popovski, “Latency analysis of systems with multiple interfaces for ultra-reliable m2m communication,” *arXiv preprint arXiv:1605.02238*, 2016.
[9] 4G Americas, “Recommendations on 5G Requirements and Solutions,” Oct. 2014.
[10] B. Holfeld, D. Wieruch, T. Wirth, L. Thiele, S. A. Ashraf, J. Huschke, I. Aktas, and J. Ansari, “Wireless Communication for Factory Automation: an opportunity for LTE and 5G systems,” *IEEE Communications Magazine*, vol. 54, no. 6, pp. 36–43, 2016.
[11] Z. Zhang, R. Hu, Y. Qian, A. Papathanassiou, and G. Wu, “D2D Communication Underlay Uplink Cellular Network with Fractional Frequency Reuse,” in *Design of Reliable Communication Networks (DRCN), 2015 11th International Conference on the*, March 2015, pp. 247–250.
[12] G. P. Fettweis, “The tactile internet: Applications and challenges,” *IEEE Vehicular Technology Magazine*, vol. 9, no. 1, pp. 64–70, March 2014.
[13] N. Panwar, S. Sharma, and A. K. Singh, “A survey on 5G: The next generation of mobile communication,” *Physical Communication*, 2015.
[14] I. da Silva, G. Mildh, A. Kaloxylos, P. Spapis, E. Buracchini, A. Trogolo, G. Zimmermann, and N. Bayer, “Impact of network slicing on 5G Radio Access Networks,” in *2016 European Conference on Networks and Communications (EuCNC)*, June 2016, pp. 153–157.
[15] NGMN Alliance, “NGMN KPIs and Deployment Scenarios for Consideration for IMT2020,” Dec 2015.
[16] 3rd Generation Partnership Project (3GPP), “TS 36.213: Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures (Release 12),” Sep. 2015.
[17] H. Shariatmadari, R. Ratasuk, S. Iraji, A. Laya, T. Taleb, R. Jäntti, and A. Ghosh, “Machine-type communications: current status and future perspectives toward 5G systems,” *IEEE Communications Magazine*, vol. 53, no. 9, pp. 10–17, September 2015.
[18] H. Min, J. Lee, S. Park, and D. Hong, “Capacity enhancement using an interference limited area for device-to-device uplink underlaying cellular networks,” *IEEE Transactions on Wireless Communications*, vol. 10, no. 12, pp. 3995–4000, December 2011.