

Title: LS on 3GPP NR Rel-16 URLLC and IIoT performance evaluation

Source: 5G Alliance for Connected Industries and Automation (5G-ACIA)

To: 3GPP TSG RAN, TSG RAN WG1

CC: 3GPP TSG SA WG1, TSG RAN WG2

Date: 26th June, 2020

Contacts: Dr. Andreas Mueller, 5G-ACIA Chairman (Andreas.Mueller21@de.bosch.com)
Dr. Afif Osseiran, 5G-ACIA Vice-Chairman (Afif.Osseiran@ericsson.com)
Alexander Bentkus, 5G-ACIA Secretary (bentkus@zvei.org)

1. Overall Description

5G-ACIA understands that 3GPP has standardized the Rel-16 RAN features related to eURLLC and IIoT in order to support industry use cases described in TS 22.104. Moreover, a new channel model for indoor factory scenario has also been developed by 3GPP RAN1 and has been published in TR 38.901. 5G-ACIA also noted that system level simulation (SLS) results were provided in TR 38.824 by multiple 3GPP partners. These simulations demonstrate the achievable performances of Rel-15 URLLC techniques with respect to the industry use cases described in TR 22.804. However, to our knowledge, limited system level performance evaluations have been conducted by 3GPP to analyse how NR Rel-16 eURLLC/IIoT technologies fulfil the performance requirements of industry use cases in TS 22.104. Specifically, 5G-ACIA kindly asks 3GPP to perform similar SLSs as described in TR 38.824 by employing Rel-16 eURLLC/IIoT features for industry use cases in TS 22.104. It is necessary for the SLSs to consider the performance requirements and influence quantities associated with the respective industry use case in TS 22.104, as well as the reasonable industry channel models in TR 38.901.

To have a better alignment on simulation assumptions with 3GPP, 5G-ACIA has identified detailed simulation parameters for use case examples pertaining to motion control (TS 22.104, clause A.2.2.1). The performance requirements of selected motion control use case are listed in Table 1.

Table 1: Service performance requirements for motion control [Table 5.2-1, TS 22.104]

Characteristic parameter				Influence quantity					
Communication service availability: target value (note 1)	Communication service reliability: mean time between failures	End-to-end latency: maximum (note 2) (note 12a)	Service bit rate: user experienced data rate (note 12a)	Message size [byte] (note 12a)	Transfer interval: target value (note 12a)	Survival time (note 12a)	UE speed (note 13)	# of UEs	Service area (note 3)
99,999 % to 99,99999 %	~ 10 years	< transfer interval value	–	50	500 μs	500 μs	≤ 75 km/h	≤ 20	50 m x 10 m x 10 m
99,9999 % to 99,999999 %	~ 10 years	< transfer interval value	–	40	1 ms	1 ms	≤ 75 km/h	≤ 50	50 m x 10 m x 10 m
99,9999 % to 99,999999 %	~ 10 years	< transfer interval value	–	20	2 ms	2 ms	≤ 75 km/h	≤ 100	50 m x 10 m x 10 m

NOTE 1: One or more retransmissions of network layer packets may take place in order to satisfy the communication service availability requirement.

NOTE 2: Unless otherwise specified, all communication includes 1 wireless link (UE to network node or network node to UE) rather than two wireless links (UE to UE).

NOTE 3: Length x width (x height).

NOTE 12: Maximum straight-line distance between UEs.

NOTE 12a: It applies to both UL and DL unless stated otherwise.

NOTE 13: It applies to both linear movement and rotation unless stated otherwise.

According to Figure 5.2-1 below (reprinted from Fig. 14 in [1]), one deployment option to implement motion control use case is to connect the controller/master (C/M) with a wire, and to connect sensors and actuators to 5G UEs. We propose to take this deployment option into account for developing simulation parameters for motion control use case examples. Our refined simulation assumptions and performance metrics are shown in Table 2. Note that simulation parameters in Table A.2.2-1(2) in TR 38.824 which are not listed in Table 2 are at the discretion of 3GPP.

Table 2: System-level simulation assumptions for motion control UC-#2

Parameters	Values	Reasons
Factory hall size	120x50m	See Section 5.2.2
Room height	10m	See service area in Table 1
Inter-BS/TRP distance	Depending on the number of TRPs, which are evenly deployed in the factory hall. Simulation company should provide the number of BSs/TRPs used in the simulation.	See Section 5.2.2
BS/TRP antenna height	1.5 m for InF-SL and InF-DL, 8m for InF-SH and InF-DH.	According to Table 1, height of factory is 10m.
Layout – BS/TRP deployment	Depending on the number of TRPs	See Section 5.2.2
Channel model	UC-2: InF-DH > InD-DL > InF-SH > InF-SL	See Section 5.2.3 and 5.2.8.
Carrier frequency and simulation bandwidth	TDD 4 GHz: 100 MHz 30 GHz: 160 MHz	There are no FDD bands identified at 4GHz. To reduce the simulation burden, it is suggested to do the performance evaluation only for TDD bands. Carrier frequency and bandwidths are inline with TR 38.824.
TDD DL-UL configuration	Simulation company should report the used DL-UL configuration.	Due to symmetric DL/UL traffic, 1:1 DL-UL configuration is recommended.
Number of UEs per service area	Up to 50 per service area, e.g., 10, 20, 40, and 50	See Table 1 and Section 5.2.8
UE distribution	All UEs randomly distributed within the respective service area.	See Table 1 and Section 5.2.4
Message size	48 bytes	See Section 5.2.8

DL traffic model	DL traffic arrival with option-1, option-2, and option-3.	See section 5.2.7
UL traffic model	UL traffic is symmetric with DL, and DL-UL traffic arrival time relationship with option-1 and option-2	See section 5.2.7
CSA requirements	UC-#2: 99.9999%	See Table 1 lower bound of CSA requirement for UC-#2 is chosen for reduced simulation burden.
Performance metrics	1) CSA: single CDF of CSA distribution of all UEs in factory hall	See Section 5.2.9 for CSA calculation with non-zero survival time;
	2) Latency: single CDF of latency distribution of all UEs in factory hall	See Section 5.1.3 for latency metric;
	3) Percentage of UEs satisfying requirements and 4) resource utilization	Metric 3) and 4) are low priority.

2. Actions

ACTION: 5G-ACIA respectfully asks

- 3GPP TSG RAN and RAN-WG1 to consider the proposed performance evaluation scenarios during the Rel-16 maintenance work as well as potential re-scoping of Rel-17 work.
- Feedback to 5G-ACIA regarding clarification of open issues and performance results/gaps from the evaluations suggested by this LS.

3. Date of Next 5G-ACIA Plenary meetings

- 15th-17th of September 2020 | conducted as e-plenary meeting
- 4th -6th of November 2020 | conducted as e-plenary meeting
- 19 – 21 January 2021 | planned as F-2-F in Stockholm
- 15– 17 March 2021 | planned as F-2-F meeting in Frankfurt

4. References

- [1] 5G-ACIA, "[5G-ACIA White Paper – Integration of Industrial Ethernet Networks with 5G Networks](#)"
- [2] Sercos Working Group, "Sercos the automation bus – Communication Specification," 1.3.1-1.12.
- [3] ETG. 1000.5 S, "EtherCAT Specification – Part 5 Application Layer service definition" v 1.0.4, 2017.

5. Annex: Information on proposed simulation parameters

The following information are taken from the internal report of ACIA work item (WI042) on performance evaluation of Rel-16 eURLLC/IIoT techniques to support industry automation uses. The WI in the following sections refers to the ACIA WI042.

5.1 Simulation metrics

5.1.1 Communication service availability

Communication service availability (CSA) is a fundamental performance metric for IIoT use cases with periodic or aperiodic deterministic traffic with strict latency requirements in TS 22.104.

TS 22.104 defines CSA as follows:

+++++

“communication service availability: percentage value of the amount of time the end-to-end communication service is delivered according to an agreed QoS, divided by the amount of time the system is expected to deliver the end-to-end service according to the specification in a specific area.

NOTE 2: The end point in "end-to-end" is assumed to be the communication service interface.

NOTE 3: The communication service is considered unavailable if it does not meet the pertinent QoS requirements. If availability is one of these requirements, the following rule applies: the system is considered unavailable if an expected message is not received within a specified time, which, at minimum, is the sum of maximum allowed end-to-end latency and survival time.

NOTE 4: This definition was taken from TS 22.261.”

+++++

It is further explained in C2.2 TS 22.104:

+++++

“Communication service availability

This parameter indicates if the communication system works as contracted ("available"/"unavailable" state). The communication system is in the "available" state as long as the availability criteria for transmitted packets are met. The service is unavailable if the packets received at the target are impaired and/or untimely (e.g. update time > stipulated maximum). If the survival time (see Table C.2.3-1) is larger than zero, consecutive impairments and/or delays are ignored until the respective time has expired."

+++++

According to Figure 5.1-1, the end-to-end (E2E) communication service interface (CSIF) for smart-manufacturing data traffic is defined between ingress and egress points of the IP layer. The end-to-end communication service interface (CSIF) for real-time process control data traffic is defined between ingress and egress points of the MAC sub-layer. It is understood that MAC sub-layer in

Figure 5.1-1 may refer to the set of aggregated protocols between IP layer and PHY layer in 5G RAN system. Depending on IIoT applications, the input data of MAC sub-layer can be IP packets or Ethernet PDUs of TSC. As a result, MAC sub-layer in Figure 5.1-1 should not be confused with MAC layer in 5G RAN protocol stack.

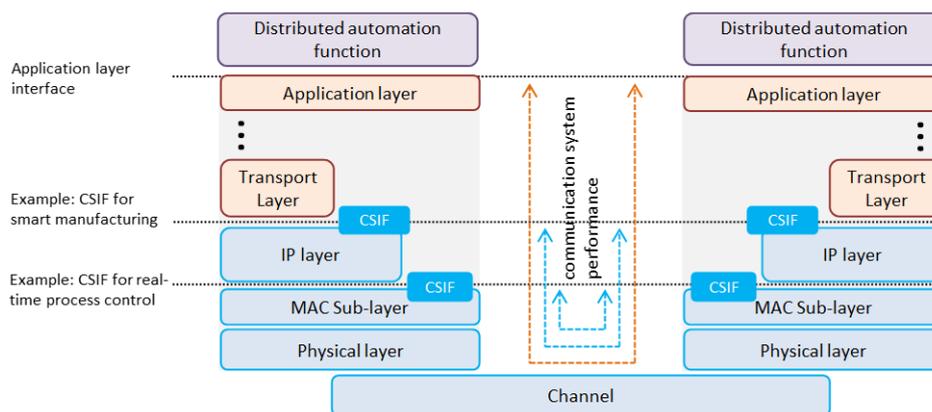


Figure 5.1-1. Performance metric at different communication service interfaces (CSIF)
 [source Fig. C.5-1 in TS 22.104]

During the study of this work item (WI), it has been further clarified that CSA in Figure 5.1-1 may be implemented as a logical communication link between a UE on the one side and a network server on the other side, or between a UE on the one side and a UE on the other side. Accordingly, the CSA performance requirements for different traffics of industry use cases in Section 5 TS 22.104 are defined for **individual logical communication links** that realise the communication services. As such, CSA is counted for a single logical link between the controller and the UE (connected to sensor/actuator). In this regard the number of connected UEs are just defining the background traffic load. CSA is not calculated as combined error probability of the communication between the controller and all of the connected UEs.

Since transport block error rate (BLER) is a typical system level simulation metric and directly affects the CSA, it is important to demonstrate how CSA can be derived from BLER or a function thereof. An example is illustrated in Figure 5.1-2.

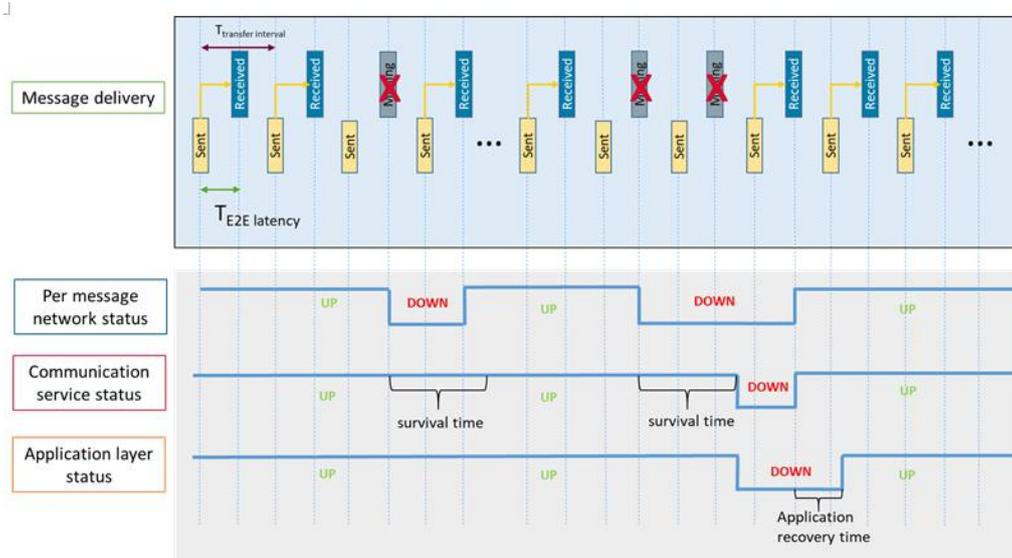


Figure 5.1-2. CSA for periodic deterministic traffic with maximum latency and nonzero survival time. [source Fig. 4.3-1 in TR 22.832]

As shown in Figure 5.1-2, transfer interval (TI) of periodic deterministic traffic is defined as T_1 , maximum latency T_L (or $T_{E2E \text{ latency}}$ in the figure), and survival time duration T_S which can be one TI or longer. Maximum latency T_L determines the time duration in which the reception of transmission (and all retransmissions, if HARQ is use,) of a transport block (TB), which encapsulates an IIoT application message, must be completed. For deterministic periodic application traffic transmission, DL semi-persistent scheduling (SPS) or UL configured grant (CG) based transmission schemes shall be applied. And to meet the stringent latency requirement, the TB allocated by configured DL/UL grant should accommodate one complete DL/UL application message to avoid fragmentation of one application message into multiple TBs transmitted over several (sub-)slots. When a message carried by the respective TB has not been correctly received at the end of latency time window, survival time window starts and lasts for time duration of T_S . At the end of survival time window, if a new message was correctly received, communication service is still deemed available (namely UP in the figure), and unavailable otherwise (namely DOWN in the figure). In the example illustrated in Figure 5.1-2, where the survival time window includes only one new message/TB transmission, in case of failed reception of two consecutive messages/TBs, the system shall be considered as unavailable for the duration equal to two TIs minus T_S (just an example). It is noted that if only one isolated message/TB is missed or incorrectly received, the communication service is still considered as available. In this case, the CSA can be derived from the probability of occurrence of two or more consecutive message/TB reception errors. Specifically, let $P_E(n)$ define the empirical probability of occurrence of exactly n consecutive message/TB reception errors, then CSA can be calculated as

$$CSA = 1 - \sum_{n=2}^{\infty} P_E(n) \frac{nT_1 - T_S}{nT_1} \quad (1)$$

In contrast to the above example, when zero survival time is required, the system shall be claimed to be unavailable if one packet is missed or incorrectly received until the maximum latency. In this case, CSA can be derived from BLER $P_E = \sum_{n=1}^{\infty} P_E(n)$ as follows.

$$CSA = 1 - P_E \tag{2}$$

As a general information, Figure 5.1-3 below shows the protocol stack of a 5G NR RAN. Several IP packets or Ethernet packets (L2 transmission) can be multiplexed in one MAC PDU Transport Block. The size of the transport block depends on the radio channel properties and is signaled or configured by DL/UL grants. Each Transport block is protected by a CRC checksum and not delivered from the PHY to the upper layer if the CRC check is missed. In this regard, one or many MAC SDU can be missed by a single MAC PDU TB reception failure. As elaborated above, CSA can be derived from the BLER when the TB can encapsulate the whole application message/packet. However when one application message is transmitted by several TBs, each of which is scheduled by the respective physical layer grant, BLER yields an upper bound for the application packet/message loss rate.

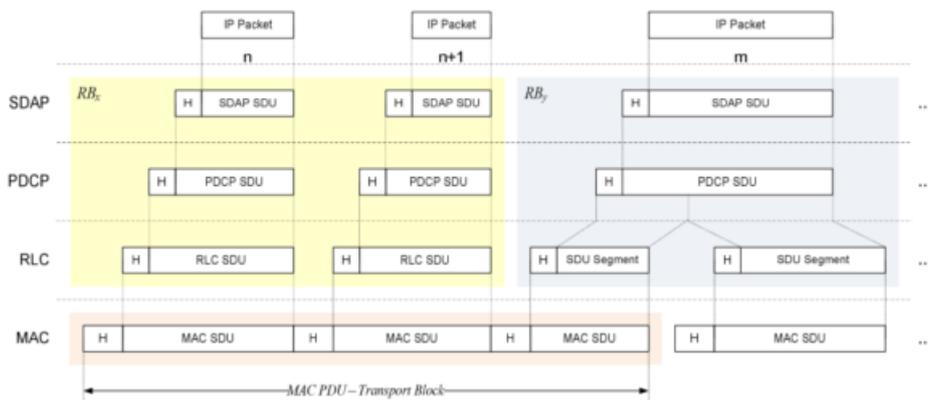


Figure 5.1-3. An exemplary packet encapsulation procedure across protocol stacks of 5G RAN

5.1.2 Communication service reliability

According to TS 22.104, communication service reliability (CSR) is defined as follows

+++++

“communication service reliability: ability of the communication service to perform as required for a given time interval, under given conditions.

NOTE 5: Given conditions would include aspects that affect reliability, such as: mode of operation, stress levels, and environmental conditions.

NOTE 6: Reliability may be quantified using appropriate measures such as meantime to failure, or the probability of no failure within a specified period of time.

+++++

CSR is further clarified in C2.2 TS 22.104 as follows

+++++

“Communication service reliability

Mean time between failures is one of the typical indicators for communication service reliability. This parameter states the mean value of how long the communication service is available before it becomes unavailable. For instance, a mean time between failures of one month indicates that a communication service runs error-free for one month on average before an error/errors make the communication service unavailable. Usually, an exponential distribution is assumed. This means, there will be several failures where the time between two subsequent errors is below the mean value (1 month in the example).

Communication service availability and communication service reliability (mean time between failures) give an indication on the time between failures and the length of the failures.”

+++++

According to the above definition, CSA and CSR are two inter-connected performance metrics, and both characterize the statistical properties of system unavailability. Specifically, CSR measures the mean time interval between two subsequent system failures, i.e., being unavailable. However, it is envisioned that CSR simulation complexity can be very extensive. For example, in many industrial use cases, CSR requires mean time interval of two subsequent failures to be ~10 years, which needs to simulate messages corresponding to 10 more years, and clearly leads to prohibited simulation complexity. Therefore, we propose to focus on CSA for availability/reliability related performance evaluation in this WI.

5.1.3 User plane communication latency

End-to-end latency is defined in TS 22.104 as follows

+++++

“end-to-end latency: the time that takes to transfer a given piece of information from a source to a destination, measured at the communication interface, from the moment it is transmitted by the source to the moment it is successfully received at the destination.

+++++

Typically, end-to-end latency in service application level is affected by both core network latency and RAN part latency. It is assumed that the CN induced latency can be negligible in this WI. As a result, this WI focuses on the latency performance of the RAN. The RAN latency performance is affected by multiple RAN system parameters, e.g., system capacity, user load, radio channel condition, etc. Given available system capacity in terms of spectrum bandwidth, and traffic load demand determined by number of users and respective data traffic needs to be served, the wireless communication system is controlled to achieve different desired performance trade-offs.

For example, when hybrid automatic retransmission request (HARQ) is applicable (e.g., if latency budget allows) and used, the transport BLER target of initial transmission can affect the average latency of user data transmission as well as the achievable system capacity in terms of number of

users to be served. Specifically, the lower the BLER target of the initial transmission is set, the smaller the likelihood for retransmission is, which further leads to a smaller average latency. However, the lower BLER target requires more radio resources provisioning for the initial transmission, and results in a smaller achievable system throughput.

When a low 5G RAN latency is achieved, the latency budget available in other parts along the end-to-end path, i.e., core network and application server part, between control applications can be positively impacted. For this reason, it is important to evaluate the cumulative distribution function (CDF) of user data communication latency in 5G RAN measured by the time delay between the time instant of the user data packet arriving at the RAN MAC input buffer at the transmitter side and the time instant of the user data packet correctly received at the RAN MAC output buffer at the receiver side.

It should be noted that control plane latency is not part of the analysis in this specific work item. Control plane latency has to be considered for the call setup, initial access, handover mechanism, etc.

5.1.4 Percentage of UEs fulfilling the requirements

For a certain use case, total amount of UEs to be served by the system is typically also given to reflect the traffic demand of the use case (UC). For a given system bandwidth and radio resource scheduling/transmission method, it is important to find out the amount of UEs who can meet the respective service requirements. This would provide valuable insight into the spectrum needs for the use case. Moreover, the percentage of UEs fulfilling the requirements is a typical performance metric of system level performance evaluation for comparing the efficiency/benefits of different scheduling and transmission schemes, therefore we also propose to include this performance metric for the evaluation. However due to the focus of this WI, this performance metric is of low priority for evaluation.

5.1.5 Resource utilization

For a use case with given spectrum and resource scheduling/transmission method, resource utilization defines the percentage of radio resources being used in the evaluation. Resource usage is also a typical performance metric to compare the efficiency of different scheduling/transmission schemes and can also provide insight for the spectrum needs for a certain UC. For some resource scheduling schemes, RU can change according to the variation of channel conditions etc. As such, we also propose to include the CDF of this metric for the performance evaluation. However due to the focus of the WI, this performance metric is of low priority for evaluation.

5.2 Simulation methodology for motion control

During the study of WI, several aspects have been recommended for simulation consideration of motion control use case. These aspects are discussed below.

5.2.1 Deployment option assumption

According to Figure 5.2-1, one possible deployment option to implement motion control use case is to have controller/master (C/M) wired connected with NR-RAN node (via UPF to NG-RAN) and integrated sensor/actuator (S/A) being mapped to 5G UEs. This seems to be a quite straightforward deployment option, we propose to take this deployment option into account for developing simulation parameters for motion control use case. Moreover, for the sake of simulation simplicity, we assume all UEs in Table 1 to be simple UEs, i.e., no IO-GW UE connecting to multiple A/S in Figure 5.2-1.

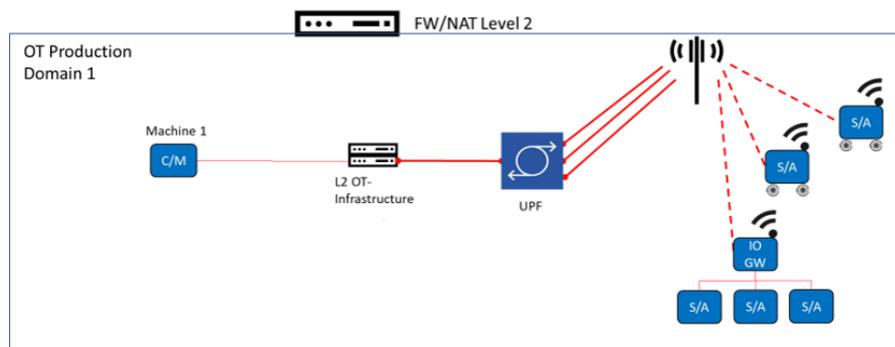


Figure 5.2-1. 5G deployment option with L2 connectivity-based motion control in one production domain [reprinted from Fig. 14 in [1]]

According to the considered deployment option in Figure 5.2-1, DL traffic refers to messages transmitted from controller to UEs, and UL traffic for messages from UEs to controller. It should be noted that in one service area of OT production factory, there can be one or several controllers which control the respective set of UEs independently.

5.2.2 Service area deployment within a factory hall

Service area is defined in TS22.104 for each of the use cases to represent a production cell in the factory hall. And channel models in TS 38.901 are defined for an entire factory hall with a size of e.g. 120x50mx10m and reflected in the factory layout for each use case. Typically, one factory hall is comprised of multiple service areas, i.e., production cells. In this WI, similar to TR38.824, we suggest to consider a factory hall of 120m x 50m x 10m, which is then fully covered by 12 service areas of 50m x 10m as shown in Figure 5.2-2.

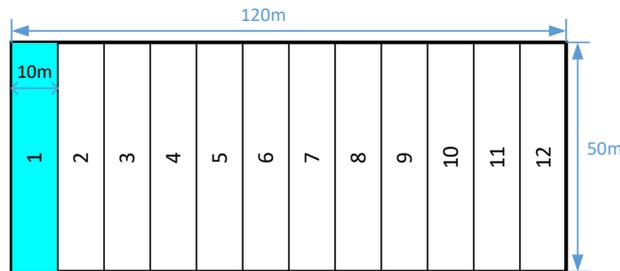


Figure 5.2-2. Service areas deployment in factory hall

5.2.3 Channel model

As described in TR 38.901, new developed 3GPP indoor factory channel models include 5 sub-scenarios, which are classified according to the factory clutter density and the antenna height of transmission reception point (TRP) of NG-RAN node. It can be overwhelming to simulate all 5 sub-scenarios for each UC. One possible prioritization of CM sub-scenarios can be based on the number of UEs in the service area. Specifically, when a large number of UEs are considered in the service area, it can be plausible to assume a large clutter density as well. Otherwise, it can be more likely to have a smaller clutter density.

5.2.4 UE moving speed model

According to note 13 in Table 1, both linear and rotation movement are considered for UE movement. However, due to lack of concrete parameter assumption, and for the sake of simulation simplicity, we propose to only assume linear movement in the simulations.

5.2.5 Unicast DL message assumption

In Table 1, each UE receives a deterministic periodic message flow. For simulation, we assume that the deterministic periodic messages are unicast messages only for the corresponding intended UEs. In NR Rel-16 Uu interface, only unicast data transmission is supported. For multicast application message transmission, as one possible implementation option, UPF can duplicate the multicast message to UE-specific PDU sessions and forward them as unicast message to the relevant NG-RAN nodes. As a result, with the assumption of unicast DL message, simulation results should also provide the radio performance for the system delivering multicast message by using unicast transmission.

5.2.6 TDD DL-UL configuration

According to deployment option in Section 5.2.1 and note 12a in Table 1, both DL and UL traffic are considered in this UC. Moreover, it is understood that the message size in Table 1 refers to individual DL and UL traffic. When TDD 5G system is simulated, DL-UL configuration used in the simulations should be reported by simulation company.

5.2.7 DL and UL traffic arrival time assumption

For DL messages for all the UEs in one service area within one transmission interval, they can be sent by controller(s) time-wise independently or in a burst manner, e.g., by summation frame in Sercos [2] and EtherCAT [3], or super-frame in PROFINET. In Sercos III, up to 127 slaves' datagrams can be transmitted in one summation frame. As such, DL traffic for all UEs can arrive at NG-RAN node in a burst or non-burst manner. To cover different typical situations, we have identified the following three options for DL traffic arrival time assumption

- Option-1: all UEs' DL messages arriving at NG-RAN node in the first transfer interval are uniformly random distributed within the TI time window.
- Option-2: all UEs' DL messages arriving at NG-RAN node in the first transfer interval are in one burst.
- Option-3: All UEs in one service area can be divided into several groups, DL messages of UEs in the same group will arrive at NG-RAN node in one burst with the following assumptions.
 - Number of groups within a service area: 2
 - Number of UEs in a group: all groups have equal number of UEs
 - 3GPP can determine to use either a pre-defined value or a random value for the burst arrival time differences between different groups.

From traffic burstiness point of view, the option-1 provides the least burst traffic while option-2 the maximum burst traffic. Radio resource scheduling algorithm needs to deal with the burst traffic properly to meet service performance requirements, e.g. latency.

For DL-UL traffic arrival time relationship, two options can be considered as follows.

- Option-1: DL and UL traffic arrival time instants are independent
- Option-2: UL traffic arrives at some pre-defined x time duration, where x can be, e.g., half of transfer interval, after the respective DL traffic arrival time.

5.2.8 Message size

In UC-#2 and #3 Table 1, small message sizes, i.e., 40 and 20 bytes, are defined. When the payload of message is smaller than 46 bytes, zero padding, i.e., adding n-padding bytes with content zeros, will be used for Ethernet frame generation assuming Ethernet Header compression. However, with bundled transmission used in summation frame described in Section 5.2.7, total payload is typically larger than the minimum payload of Ethernet frame, so no zero-padding will be used in this case. To reduce the simulation burden, and take into account the message size constraint, we propose to only focus on UC-#2 with modified message size of 48 bytes.

5.2.9 CSA calculation

As described in Section 5.1, CSA should be calculated on an individual logical link basis, i.e. per-UE basis in the simulations. When each TB can encapsulate a complete application message, the CSA calculation can be obtained from BLER by (1) and (2) for non-zero and zero survival time, respectively.