

Agenda item: 7.2.9.2
Source: Qualcomm Incorporated
Title: Potential Techniques for UE Power Saving
Document for: Discussion/Decision

1 Introduction

In RAN plenary #82, a draft TR 38.804 for UE power saving was endorsed [10]. In this contribution, the sections are structured in a similar way as in the draft TR in terms of how the potential techniques are categorized. Please refer directly to the draft TR for description pertaining to the potential techniques, many of which are the subjects of discussion in the following sections.

There are two major sections in this contribution. Section 2 contains mostly new material on top of R1-1900912 from last meeting, and Section 3 is mostly a resubmission of R1-1900911 with some minor refresh on the observations/proposals.

2 Power Saving Signal/Channel for Triggering Adaptation to UE Power Consumption

2.1 PDCCH-based Wake Up Signaling

2.1.1 Design principles

In Figure 1, the workflow of PDCCH-based WUS (abbreviated as “PDCCH-WUS” hereafter) is illustrated. At first, UE wakes up from sleep (i.e., C-DRX off periods) and tries PDCCH-WUS decoding. When PDCCH-WUS is not decoded in the 1st wake-up (i.e., WUS occasion), the UE can go back to sleep immediately, skipping the 2nd stage wake-up (i.e., ON duration); otherwise, UE continues with the 2nd stage wake-up and to the ON duration for DL assignment and data reception.

It may sound counter-intuitive that PDCCH-WUS can achieve power saving. If the ON duration can be configured to the same duration as the WUS occasion (e.g., 1 slot), the UE would perform the same workload of PDCCH decoding regardless of whether this is for WUS or for data scheduling. If the UE does not decode any PDCCH, the end result is the same, that is, the UE goes back to sleep, i.e., C-DRX off periods. One may ask, why would there be power saving with PDCCH-WUS?

In reality, UE’s power consumption for ON duration and for WUS occasion can be very different, even if the signaling scheme for both is based on PDCCH. During WUS occasion, the UE does not expect to receive a same-slot grant for PDSCH or have to be ready to transmit PUCCH in response to PDSCH reception. Also, once the UE decodes PDCCH-WUS in the WUS occasion, there is a time offset to the ON duration. Therefore, during WUS occasion, a low power implementation can be achieved by optimizing, at least: (i) the PDCCH processing timeline, (ii) the amount of hardware that needs to be brought online, (iii) the operating point in terms of the voltage levels and clock frequencies of the hardware, and (iv) potentially the Rx bandwidth and the number of Rx antennas for PDCCH-WUS. Some aspects of the optimization are discussed more in detail in the Appendix of [2] and our companion paper [15]. Only when a PDCCH-WUS is decoded, the UE performs wake-up of additional hardware and processing (e.g., by performing BWP switching, CSI-RS processing, etc.) to get ready for potential DL/UL scheduling. This 2nd stage of wake-up consumes additional energy but can be skipped for empty C-DRX cycles.

Observation 1: Two-stage wake-up is the key idea for PDCCH-based WUS to facilitate power saving for DRX.

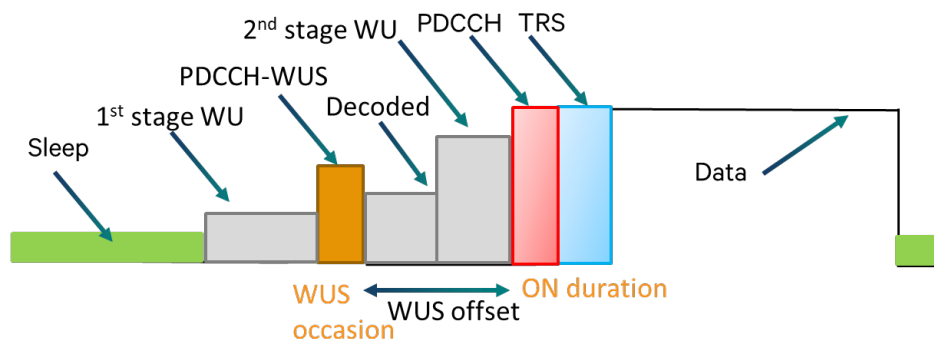


Figure 1: Two-stage wake-up timeline for PDCCH-based WUS

In our past contribution [1], two strategies were considered for the introduction of PDCCH-WUS in NR. The first one is re-using and extending the BWP adaptation framework of Rel-15, and the second is designing a special-purpose PDCCH-WUS. Although the first strategy may help minimize specification work across RAN1 and RAN2, the second strategy may lead to further optimized design.

Observation 2: PDCCH can be studied as a candidate WUS waveform in Rel-16. Two strategies of introducing PDCCH-WUS in Rel-16 can be considered:

- 1) **Enhancement of BWP adaptation framework of Rel-15**
 - a. **Special configuration for the starting BWP (or “wake-up BWP”) for C-DRX ON duration**
 - b. **Minimum k0/k2 threshold configuration for the starting BWP**
- 2) **Design of new dedicated PDCCH-WUS**
 - a. **UE-specific WUS occasion and offset configuration**
 - b. **PDCCH structure, CORESET/search space, BWP configuration specific for the WUS occasion**
 - c. **DCI fields for PDCCH-WUS that help UE prepare for the transition to C-DRX ON duration**

Proposal 1: For the study of PDCCH as a candidate WUS waveform, designing a new dedicated PDCCH-WUS leveraging Rel-15 PDCCH should be the preferred approach.

2.1.2 Multiplexing of PDCCH-WUS

Resource sharing of the PDCCH-WUS with other signals, including other non-DRX or active UE’s physical channels/signals and other UE’s WUS, should be the first-order design topic of PDCCH-WUS in Rel-16. A key factor of the design is maintaining backward compatibility with Rel-15. Consequently, we may mostly rely on the existing configuration parameters and well-defined multiplexing rules for PDCCH. For instance, a group of UEs may be configured with a specific C-DRX cycle/offset and WUS occasion that are distinct from other groups of UEs. Furthermore, among the UEs in the same group, each UE can be configured with a specific BWP, CORESET/search space, scrambling ID, RNTI, etc., to monitor and distinguish its own PDCCH-WUS from others.

Observation 3: Rel-15 PDCCH configuration and multiplexing rules can be leveraged for PDCCH-WUS.

Although the existing Rel-15 PDCCH framework may render enough flexibility for PDCCH-WUS resource configuration, there are some cases that requires special attention. For example;

- In a heavily loaded scenario, a large number of UEs may share the same WUS resources, such as a WUS occasion, BWP, and CORESET. Due to the high contention for the resources, some UE’s PDCCH-WUS may be blocked and the scheduling for the UE can be delayed to the later C-DRX cycles, which results in impaired latency and user-experience.
- In Rel-15, the first three symbols within a slot are predominantly used for control channels. If a WUS occasion for PDCCH-WUS shares the same resources as the control resources for other active UEs, it may worsen the control congestion in a heavily loaded scenario.
- In mmW (e.g., FR2) systems with analog beamforming, TDM-based multiplexing of WUS resources may be preferred across spatially non-co-located UEs. Therefore, sharing the same WUS occasion among UEs may be restrictive and the overhead for WUS resources may further increase with a large number of UEs.

From the above argument, we may conclude that, for PDCCH-WUS configuration, further enhancement is needed on top of the Rel-15 PDCCH configuration. A potential enhancement is allowing the search space of PDCCH-WUS, i.e., the WUS occasion, to span any consecutive OFDM symbols of a slot, so that a WUS occasion can be separated from other WUS occasions or ordinary PDCCH search spaces of other UEs. An example of this configuration is shown in Figure 2.

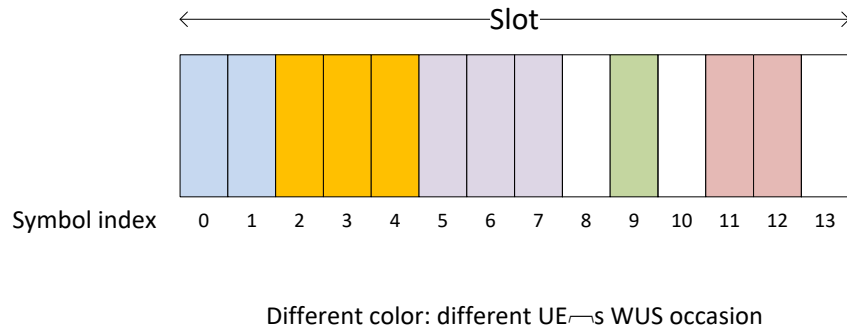


Figure 2: Configuration of WUS occasions

Proposal 2: Flexible time-domain location of a search space can be considered as a Rel-16 enhancement for PDCCH-WUS configuration.

We may further note that, based on the agreement in RAN1#91 (PDCCH monitoring Case 1-2 or Case 2), the flexible time location of a search space is already supported in Rel-15 (Section 10.1 in TS 38.213), at least for 15kHz numerology. Therefore, the additional specification effort could be marginal.

2.1.2.1 Multi-UE PDCCH-WUS

A group-common PDCCH signal can be used as a WUS for multiple UEs. A group of UEs can be configured to monitor the same GC PDCCH. The UEs interpret the presence of the GF PDCCH as a wake-up signal or the DCI content can further indicate which UEs in the group are the intended target of the WUS. In case of DCI-based indication, the maximum number of UEs supported by this signaling mechanism is proportional to the DCI size. Section 2.1.4 shows the performance of decoding PDCCH for different DCI sizes.

2.1.3 PDCCH-WUS Resource Overhead

The amount of resources occupied by PDCCH-WUS corresponds to the PDCCH aggregation level, which is set by the gNB according to the expected signal quality at the UE. It should be noted that PDCCH-WUS can carry information in DCI beyond simply waking up the UE. Therefore, the resources utilized by PDCCH-WUS can carry more than the equivalent of a single bit of information.

Observation 4: PDCCH-WUS resource utilization per UE or UE group is determined by the aggregation level set by the gNB.

Since the DCI size for PDCCH-WUS can be small as will be discussed in Section 2.1.5, the performance will be improved compared to sizes typically used by DCIs scheduling a transmission. Therefore, lower aggregation level can be used for PDCCH-WUS compared to PDCCH used for scheduling, reducing its resource overhead.

2.1.4 PDCCH-WUS Performance

The minimum DCI size is 12 bits, ensured by padding, in addition to a 24-bit CRC, resulting in a minimum size of 36 bits. Figure 3 shows the SNR required to reach 10^{-3} misdetection rate for three DCI sizes (including the 24-bit CRC): 36, 44, and 54. Performance results are provided for different aggregation levels and number of receiver antennas. It can be observed that PDCCH-WUS provides good mis-detection rate, especially with the smaller DCI sizes. Performance can be further improved by using aggregation-level 16. These simulation results assume that DCI carries a payload, the UE detects whether it is the intended recipient based on the RNTI used to scramble the CRC (standard PDCCH reception procedure).

The false alarm rate of PDCCH was extensively studied and was found to be less than 1.5×10^{-21} when a polar decoder with list size of 8 was used (as was used for the performance results in Figure 3) [3]. Since the false alarm rate is well below the 10% target, some of the distributed CRC bits can be used to aid in the polar decoding processing, improving its performance.

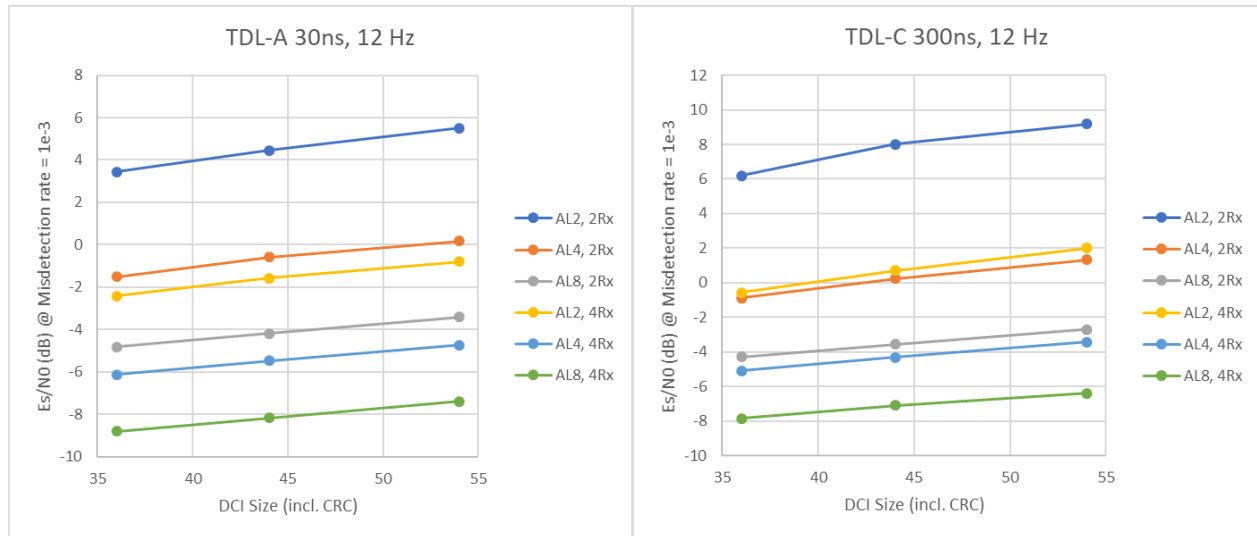


Figure 3 SNR required to reach PDCCH misdetection rate of 10^{-3} for different DCI sizes with realistic channel estimation.

Observation 5: PDCCH-WUS meets the misdetection rate criterion and provides extremely low false alarm rate.

The UE can also perform sequence detection on the PDCCH DMRS in lieu of or prior to decoding PDCCH as is possible with a CSI-RS-based or sequence-based WUS. The PDCCH DMRS sequence is generated using a configurable seed, which can further aid the UE in determining whether it is the intended recipient of the PDCCH-WUS and deciding whether to decode DCI.

2.1.5 Potential Payload Information

One advantage of PDCCH-WUS is that DCI can carry multiple bits of information that indicate more than wake-up to the UE. For example, the wake-up DCI can contain a BWP ID for the UE to switch to, providing additional flexibility to the scheduler to select a desired BWP and enabling an earlier BWP switch compared to waiting until the ON duration to signal the new BWP. Aperiodic CSI trigger is another field that could be included in the wake-up DCI, also reducing the latency of reporting CSI compared to waiting until the ON duration.

In the case of group WUS, the wake-up DCI can also include information to indicate whether a specific UE should wake-up or not. This indication can be either a direct bit map or a hash-based scheme for example.

Observation 6: DCI in PDCCH-WUS can carry information in addition to a binary wake-up indication. For example, a target BWP ID, A-CSI trigger, and a fine-granularity UE wake-up indication.

2.1.6 Complexity of Detecting PDCCH-WUS

Direct decoding of PDCCH-WUS comprises channel estimation, descrambling, and polar decoding. The UE can also detect the presence of the DMRS sequence first using correlation methods and then only perform channel estimation and decoding based on the sequence detection outcome. If DMRS sequences are generated using different seeds for different UEs or UE groups, this can further aid the UE in determining whether it is the intended recipient prior to performing PDCCH decoding.

Observation 7: The UE can use the DMRS sequence to detect PDCCH-WUS presence prior to decoding DCI, reducing complexity and avoiding unnecessary decoding attempts.

2.2 CSI-RS-based Wake Up Signaling

2.2.1 Design principles

In our past paper [2], we propose the use of pre-wake-up (PWU) windows, during which UE can measure CSI-RS resources. PWU windows provides opportunities for UEs in the C-DRX mode to prepare for the upcoming ON duration. Therefore, the UE can less rely on SS/PBCH blocks, which requires additional wake-up during C-DRX off periods, and the power consumption can be reduced.

In order to reconcile the concept of the PWU window with WUS, we first need to answer to a question: is the preparation stage needed for every ON duration? In our view, the answer is negative; if there is no imminent data for the UE in a C-DRX cycle, the corresponding ON duration is empty and the preparation stage may consume unnecessary UE power and network resources. Therefore, a potential enhancement of the PWU concept is to transmit CSI-RS only when it is needed, i.e., when the UE need to wake up for the next ON duration. Then the UE will first detect the presence of CSI-RS within the PWU window and decide whether to skip the next ON duration or not. Once the UE detects the presence of CSI-RS resources, it can use the same resources to warm-up for the upcoming ON duration. Notably, this naturally aligns with the concept of WUS and establishes the baseline of CSI-RS-based WUS (abbreviated as “CSI-RS-WUS” hereafter).

Observation 8: CSI-RS-type waveform can perform as both WUS and warming-up resources.

To embed information payload (such as DCI of PDCCH-WUS) into CSI-RS-WUS, different resources and scrambling sequences can be mapped to the WUS based on the payload, and the UE may perform blind decoding of CSI-RS-WUS with different hypotheses. However, compared to PDCCH-WUS, carrying a relatively large payload (e.g., 12bits or larger) may not be feasible for CSI-RS-WUS.

Proposal 3: CSI-RS can be studied as a candidate WUS waveform in Rel-16.

2.2.2 Multiplexing of CSI-RS-WUS

As for resource sharing of CSI-RS-WUS, the same discussion as PDCCH-WUS in Section 2.1.2 can be repeated, noting that Rel-15 already offers a rich platform of CSI-RS configuration and it can be reused for CSI-RS-WUS. For example, the resources for CSI-RS-WUS may be associated with different FDM/CDM patterns and scrambling IDs and, also, multiple WUS occasions can be staggered in time with different C-DRX offset values. At the same time, for the other non-DRX or active UEs, RateMatchPatterns or ZP-CSI-RS resources may be configured and dynamically indicated for rate matching around the CSI-RS-WUS resources.

Observation 9: Rel-15 CSI-RS configuration and multiplexing rules can be leveraged for CSI-RS-WUS.

To analyze the multiplexing capability and resource overhead of CSI-RS-WUS, we assume configuration parameters in the following table.

Configuration	Values
Number of CSI-RS symbols	4 in total, 1 per UE
Frequency density	1 REs/RB/port per UE
Number of RBs	48 in total, 24 per UE
Min WUS occasion periodicity	20ms
CC BW	64RBs (assuming 100MHz FR2)

Based on the assumed configuration parameters, the resource overheads of CSI-RS-WUS accommodating different number of UEs are derived in the following table.

#WUS_UEs	CDRX cycle		
	40ms	160ms	320ms
96	0.067%	0.017%	0.0084%
192	0.13%	0.033%	0.017%
384	-	0.067%	0.033%

768	-	0.13%	0.067%
1536	-	-	0.13%

2.3 Further Issues in Wake Up Signal Design

2.3.1 Beam Sweeping Wake Up Signal in FR2

2.3.1.1 Motivation of WUS beam sweeping

In our past contribution [1], we pointed out that the link quality degradation during C-DRX off periods can be more desperate in FR2 due to beam deviation. There can be many causes of beam deviation, such as UE rotation, UE mobility or beam blocking, etc. During the active time, the beam deviation can be compensated by a beam management process. That is, as a part of the beam management process, the UE can persistently measure the quality of beam paired links, e.g., using SSBs and CSI-RS, and report it to the gNB. As a response, gNB can re-assign or switch beams for the UE. However, due to lack of active beam management during C-DRX off periods, the link quality may decay severely. The direct result of this is the poor detection performance of WUS, which has a critical impact on the WUS scheme.

To combat the beam degradation issue, beam sweeping for WUS can be considered for improved beam diversity. The gNB and UE can be coordinated to use a specific set of WUS resources associated with different beams and, if the UE detect the WUS on at least one of the resources, the UE can wake up.

Proposal 4: At least in FR2, beam sweeping can be applied for WUS transmission.

Note that beam sweeping is not limited to a specific WUS waveform but can be fitted with both PDCCH-WUS and CSI-RS-WUS. In fact, based on Rel-15 configuration framework, CSI-RS-WUS is more beam-sweeping-friendly, as beam-sweeping is already a part of CSI-RS-based beam management. Thus, considering the required specification effort, CSI-RS-WUS may be suitable for FR2.

Observation 10: For FR2, CSI-RS-WUS with beam sweeping may be more well-suited than PDCCH-WUS.

However, even with all additional specification efforts, PDCCH-WUS has advantages over CSI-RS-WUS, such as high detection reliability (with the help of CRC) and a capability of large payload delivery. Therefore, PDCCH-WUS should not be scoped out for FR2 WUS design.

2.3.1.2 Link-level performance of beam-swept wake-up signals

In this and following sections, we investigate the link-level performance of WUS in FR2, both for PDCCH-WUS and CSI-RS-WUS. In particular, we focus on the evaluation of WUS beam sweeping in the presence of beam deviation.

The goal is to quantify the relationship between the WUS misdetection probability and the beam deviation during a C-DRX off period. In order to model the spatial/angular-domain channel variation, we assume that the UE's orientation changes at a constant rate, e.g., 180 degrees per second. The channel variation model is illustrated in Figure 4. Initially, the UE has a uniformly random bearing angle, while the downtilt and slant angles are fixed to zero, and a set of N best beam pairs in terms of L1 RSRP is found. Then, the UE goes to sleep during a certain length of C-DRX off period and wakes up to detect WUS. For simplicity, we further assume that the length of C-DRX off period is equal to the length of the C-DRX cycle (i.e., the length of ON duration and WUS occasion is ignored), and there is no additional (e.g., SSB-based) beam refinement during the C-DRX off period. Therefore, the total angular deviation of the channel is determined as

$$\text{Angular deviation: } \Delta\alpha = 180 \cdot \Delta t \text{ [deg]},$$

where Δt is length of the C-DRX cycle in seconds.

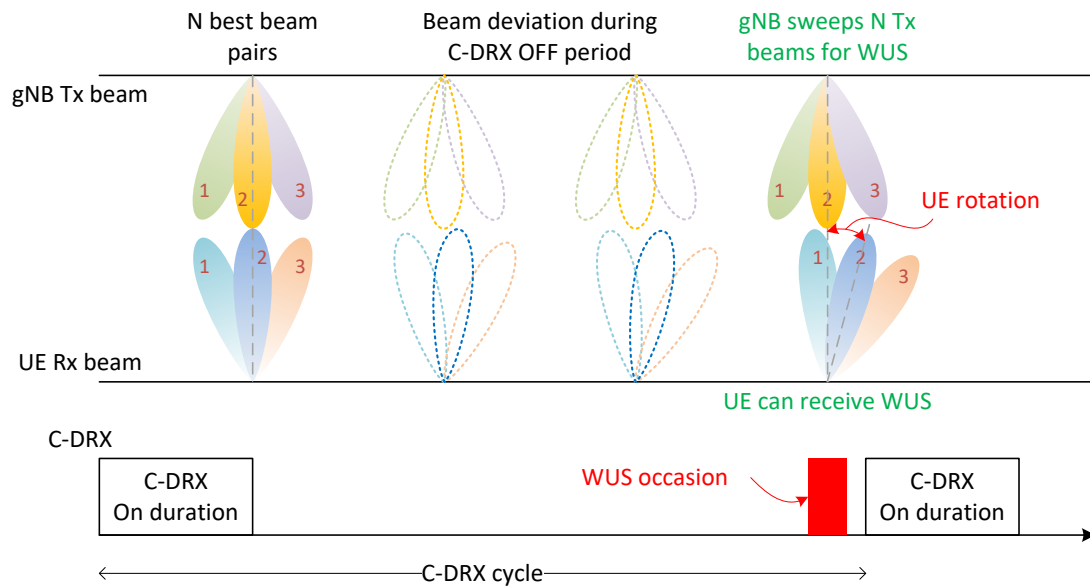


Figure 4: Beam deviation during C-DRX operation

As captured in [10], the link-level simulation parameters specified in Table A1.5-1 in TR38.802 are assumed. Additional information for PDCCH-WUS configuration are shown in the following table.

Setting (PDCCH-WUS)	Configuration
PDCCH payload size	48 bits (including CRC)
CORESET size	48 RBs & 1 symbol
CCE aggregation level	4 or 8
CCE-to-REG mapping	Non-interleaved

For CSI-RS-WUS, the configuration in the following table is assumed.

Setting (CSI-RS-WUS)	Configuration
Number of ports	1
Density	3 REs/RB/port
Bandwidth	24RBs or 48 RBs
Power boosting	0dB
False-alarm probability	1%

Also, the antenna array and analog-beamforming-related parameters are listed in the following table.

Setting (FR2)	Configuration
Carrier frequency	30GHz
Channel model	CDL-C with delay scaling value of 300ns
gNB antenna array	{M, N, P} = {64, 4, 2}, single panel

gNB analog BF codebook	2D DFT codebook with 256 beams (64 [azimuth] * 4 [elevation])
UE antenna array	{M, N, P} = {4, 2, 2}, two panels with 180° facing angle
UE analog BF codebook	2D DFT codebook with 8 beams
UE rotation speed	180° per second

2.3.1.3 Evaluation results for PDCCH-WUS

In Figure 5 and Figure 6, the misdetection performance of PDCCH-WUS is presented for different numbers of sweeping beams. The SINR values of interest are 3dB and -6dB for Figure 5 and Figure 6, respectively, where the SINR is defined for the post-beamforming signal with the best beam pair in terms of L1 RSRP.

Setting the target misdetection probability to 0.1%, in Figure 5 (SINR=3dB), the target performance can be achieved for 40ms C-DRX cycle without any beam sweeping. However, with 160ms C-DRX cycle, it is observed that beam sweeping is inevitable. For example, with AL8, using 3 or more beams to sweep for PDCCH-WUS can achieve the target performance with 160ms C-DRX cycle. Also, with AL4, the target performance can still be attained for 160ms C-DRX cycle, with a larger number of beams (e.g. 7 beams) than AL8. However, for a longer C-DRX cycle, such as 320ms, the target performance cannot be attained for all cases. As a consequence, in rapid beam deviation scenarios, the C-DRX cycle should be chosen to be short enough or additional beam management during C-DRX off period should be considered.

In Figure 6 (SINR=-6dB), even the baseline performance (w/o beam deviation) cannot achieve the target performance and the gain of WUS beam sweeping is not advocated. In principle, at a very low SINR regime, despite additional efforts and resource requirements for adopting PDCCH-WUS, only marginal return would be expected. In such a case, the UE may choose to operate in a legacy C-DRX mode (Rel-15 C-DRX operation).

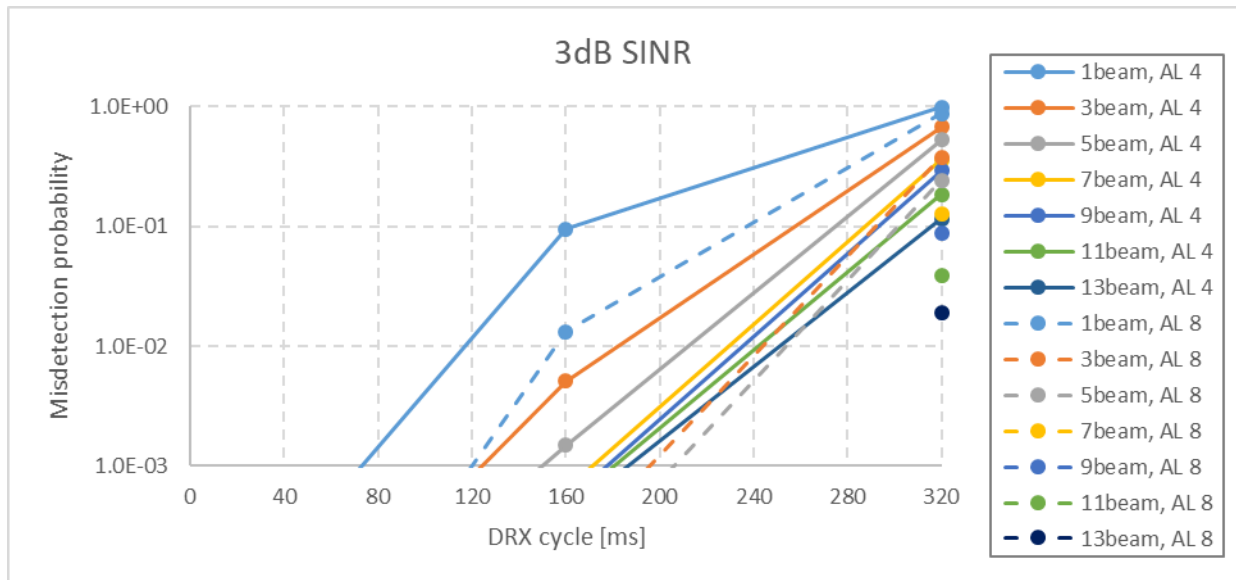


Figure 5: PDCCH-WUS misdetection probability at SINR=3dB

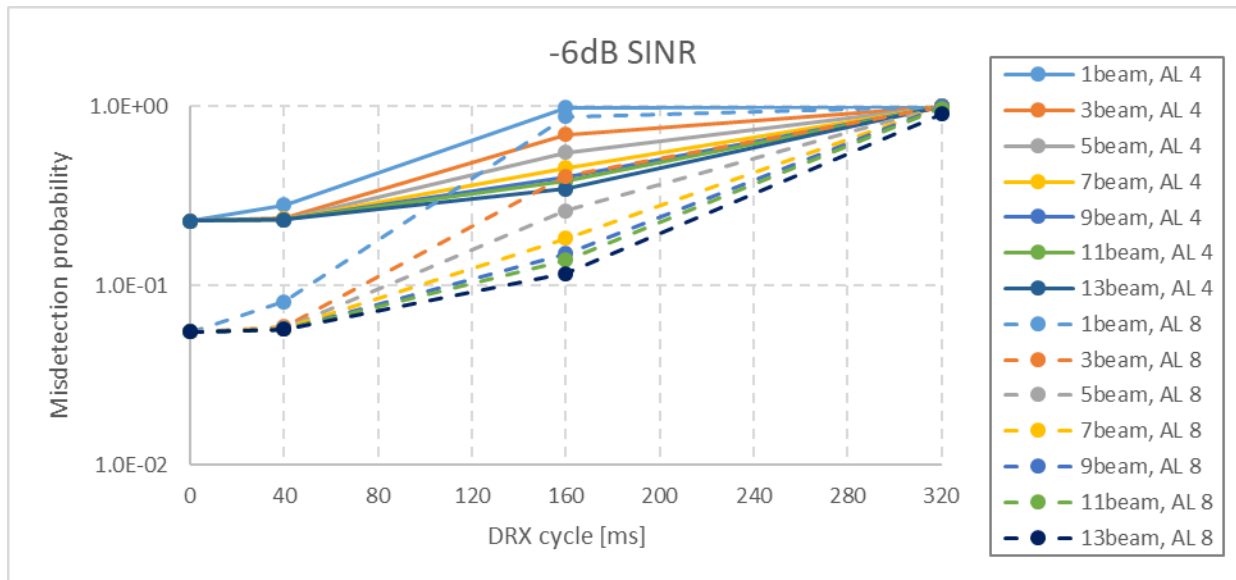


Figure 6: PDCCH-WUS misdetection probability at SINR=-6dB

2.3.1.4 Evaluation results for CSI-RS-WUS

In Figure 7 and Figure 8, the misdetection performance of CSI-RS-WUS is presented for SINR values of 3dB and -6dB, respectively. Overall, a similar trend to the case of PDCCH-WUS is observed for CSI-RS-WUS, and we can conclude that beam sweeping is quite effective for alleviating the WUS beam degradation issue.

Comparing the performance of PDCCH-WUS and CSI-RS-WUS, it is seen that CSI-RS-WUS has consistently lower misdetection probability than PDCCH-WUS with the same number of beams. However, this is mainly due to the fact that PDCCH-WUS can carry a larger data payload than CSI-RS-WUS and does not necessarily imply domination of CSI-RS-WUS.

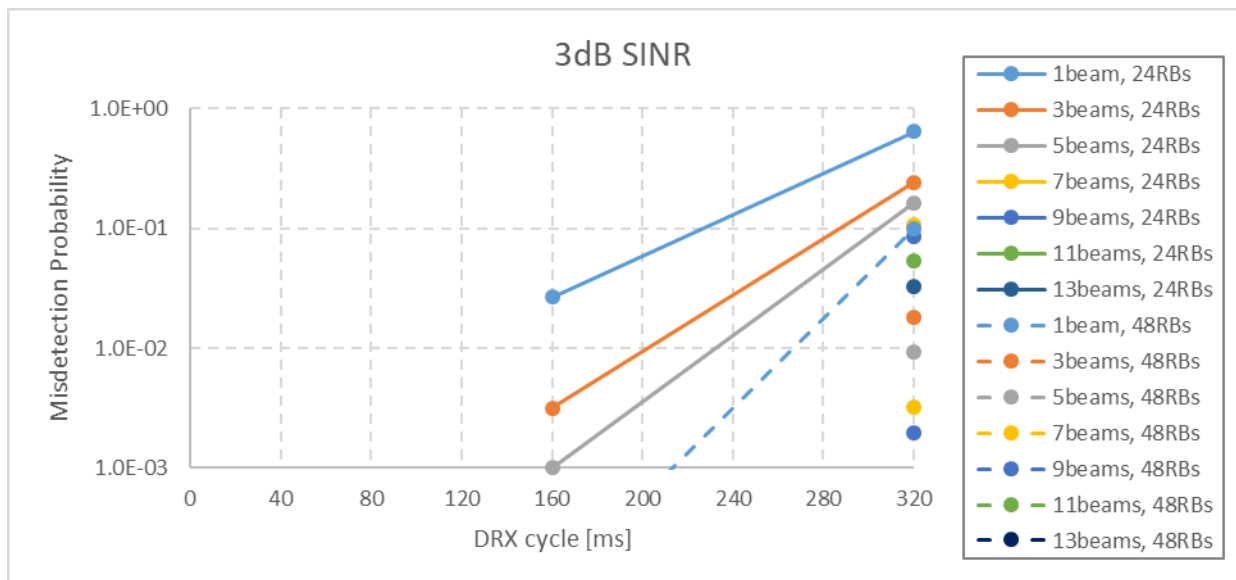


Figure 7: CSI-RS-WUS misdetection probability at SINR=3dB

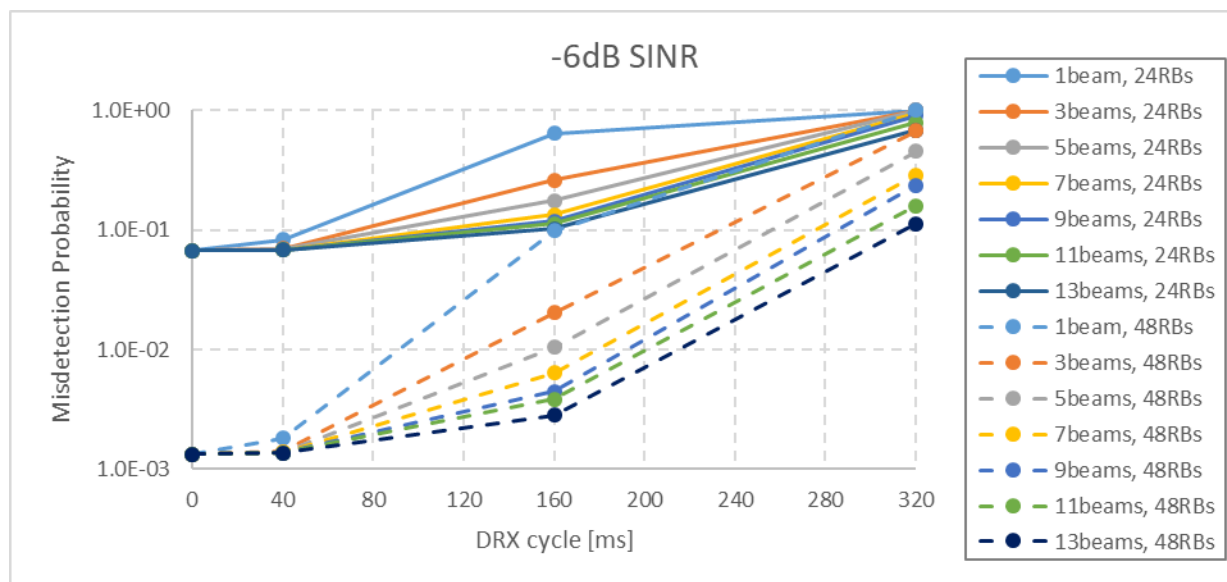


Figure 8: CSI-RS-WUS misdetection probability at SINR=-6dB

2.3.2 Incremental Power Saving for Other WUS Schemes

If PDCCH-WUS or CSI-RS-based WUS can already deliver two-third reduction in awake energy compared to normalized (i.e. 1-slot) of ON duration (see discussion of this assumption in [15]), how much incentive/benefit is there to consider other WUS schemes, e.g. sequence-based, which may be able to deliver 90% awake energy reduction?

We have a preliminary analysis as follows. Let's evaluate the impact on C-DRX average power for the case of 80 msec C-DRX cycle. And we assume empty C-DRX use case contributes to 10% of overall DoU.

Let's assume chipset sleep current is 1 unit (similar to deep sleep).

	Baseline	PDCCH/CSI-based WUR	Other WUR
Awake energy	170	57 (33% of baseline)	17 (10% of baseline)
Average "empty" CDRX power	3.1	1.7 (55% of baseline)	1.2 (39% of baseline)
DoU contribution	10%	5.7%	4.2%

As one can see, the improvement in terms of DoU contribution is reduction from 5.7% to 4.2%.

Even if WUR awake energy can be optimized to zero, average CDRX power would be floored at 1 (sleep current), DoU contribution still hovers at 3.4%.

Above is a simple analysis assuming all C-DRX cycles are empty, for which WUS should offer the highest gain. In [15], simulation with realistic traffic models is used to evaluate the difference between WUR with 1/3 vs 1/10 of awake energy with respect to the baseline. The former is the power model proposed for PDCCH/CSI-based WUR, and the latter is for another WUS/WUR scheme. The table capturing the power consumption for different cases with WUS is repeated below:

	Web browsing /w DRX (160, 100)	FTP /w DRX (160, 100)	FTP /w DRX (40, 10)	IM /w DRX (320, 80)

WUR (~1/3 energy)	5.6	39.7	19.8	7.1
WUR (~1/10 energy)	5.0 (11% less)	39.0 (1.7% less)	16.8 (15% less)	6.8 (4.2% less)

It can be seen that for a three-fold reduction in WUR energy, the resultant gain in overall average power for the use case is quite low (at most 15%). This shows that while wake-up receiver energy consumption should be low, it does not need to be over-optimized as the marginal gain would diminish beyond certain point.

Observation 11: If PDCCH-based or CSI-RS-based WUS can deliver significant awake energy reduction for detection compared to normalized baseline of ON duration, the gain for further overall power saving may diminish even if another WUS scheme can achieve much lower awake energy for detection.

2.4 Signaling to Trigger Power Saving during Active Time

In our past paper [2], we show our views on power saving schemes for connected mode UEs operating in the active time. A power saving technique during the active time may be triggered implicitly by an event, or explicitly by a power saving signal. In general, a triggering scheme is tightly coupled with a specific power saving technique and thus they cannot be discussed separately. Also, due to the wide variety of power saving techniques in the application scenarios, requirements, and benefits, it would not be possible to have a unified trigger signaling in all cases. For example, for some time-domain power adaptation techniques, PDCCH-based trigger signaling may be the best option because the UE would keep monitoring control channels somehow during the active time. In some other cases, the power would be adapted in semi-static manner based on UE assistance information, via RRC or MAC-CE signaling.

Observation 12: Discussion on signaling that triggers a power adaptation technique should be coupled with the discussion on the power adaptation technique. A unified signaling scheme may be undesirable.

3 UE Adaptation to the Traffic and Power Consumption Characteristic

3.1 Adaptation to the variation in frequency

3.1.1 Power Saving for CA

3.1.1.1 Motivation for SCell power saving

Generally, carrier aggregation can be power efficient if there is enough data being served on the carriers. The problem area is during the periods when utilization of SCells is low, DL control monitoring on SCells can be costly in terms of power consumption. It would be more power efficient to suspend activities on SCells (including PDCCH monitoring) when the traffic activity is low. In LTE and NR Rel-15, SCell activation and deactivation was designed for such power saving purpose. However, in practice, because the latency for activating a SCell is quite high (24~36 milliseconds for LTE; can be higher for Rel-15 NR), network tends to activate SCells and leave them activated even during periods of low data activity and when the SCells are not used. Obviously, if NR can greatly reduce SCell activation /deactivation time, SCell can be deactivated and activated more frequently with less overhead, and power efficiency for carrier aggregation can improve.

3.1.1.2 SCell Dormancy

In Rel-15 LTE eMTC [9], a new SCell state, distinct from legacy activated and deactivated state, is introduced. This new state allows periodic CSI of the SCell to be measured and reported, but otherwise is similar to legacy deactivated state: no data transfer and no PDCCH monitoring is expected on the SCell. The advantage is that the latency of transition into active state from this new state is reduced significantly compared to transition from deactivated state. The UE power consumption in the new state is much lower than that in active state and can be slightly higher than that in deactivated state. The ability to do fast state transition between low power new state to high power active state results in reduced overall UE power consumption for most traffic patterns.

In NR, a similar approach can be considered. An intermediate state between SCell activation and deactivation, where UE does not PDCCH monitoring on the SCell and does not expect to be scheduled on the SCell. Other periodic activities on the SCell including CSI measurement and reporting, beam management, RRM may still be performed at much reduced duty cycle, but at a level to allow transition to activation state can be very quick. It can be viewed as a way to “suspend” SCell operation for moderate level of power saving.

Observation 13: Not monitoring PDCCH on SCell is a necessary but not a sufficient condition for improving SCell energy efficiency. Other periodic activities including CSI measurement and reporting, beam management, and/or RRM may still be performed at much reduced duty cycle. There is tradeoff between minimizing SCell power consumption level and maintaining sufficient state information about the SCell such that UE can very quickly transition the SCell to activation state when traffic arrives.

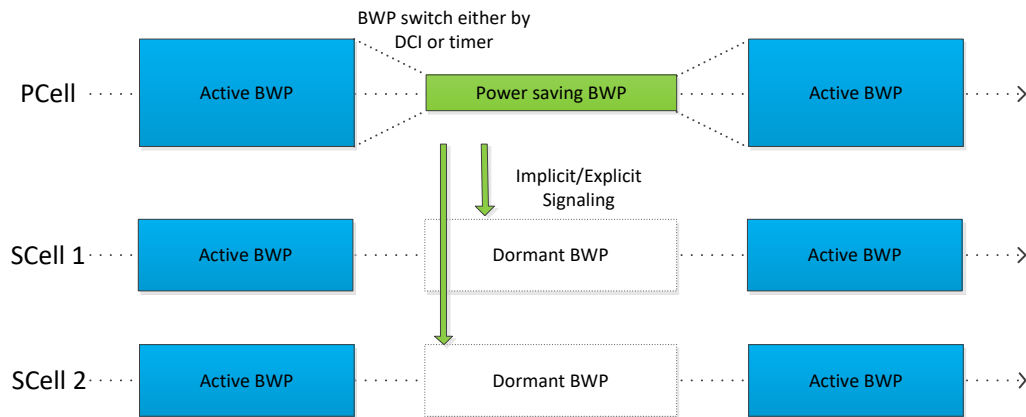
The latency for transition in and out of dormancy state should be smaller than the latency for SCell activation/deactivation. In NR, BWP adaptation is supported with BWP transition latency up to the order of 2 milliseconds. It makes sense to consider a scheme that leverages BWP adaptation framework for SCell dormancy support, as discussed in the next section.

3.1.1.2.1 Dormant BWP for SCell

From the perspective of BWP adaptation applied to SCell, rather than adapting the bandwidth of each individual SCell, being able to put SCell into power saving state and bringing it back briskly would be more effective for power saving. SCell dormancy state can correspond to a designated BWP with special attributes supporting low power operation. Such BWP can be called “dormant BWP” where no uplink or downlink grant is expected by the UE, but other periodic activities such as CSI measurement and reporting may happen at a much larger periodicity.

The details of how “dormant BWP” can be configured can be discussed further. During period of low traffic activity, activated SCell(s) can have its active BWP switched to the dormant BWP; The UE would continue monitoring DL control channel on the PCell, but not on the SCell(s) with dormant BWP being active. The actual signaling mechanism to put SCell in and out of dormant BWP can be done via specialized BWP DCI signaling on the PCell. Also, the default BWP for the SCell can be configured to be the dormant BWP, so that when BWP timer expires, the SCell autonomously switches to the dormant BWP for power saving.

Alternatively, a more simplified scheme is to base the switching of SCell’s dormant BWP on PCell’s active BWP status. For example, one BWP on PCell can be designated as the “power saving BWP” and whenever it is active, SCell’s active BWP is implicitly switched to the dormant BWP (illustrated below). It is most logical to configure the power saving BWP the same as the default BWP on PCell, so that upon BWP timer expiry, SCells are switched to the dormant BWP. For better flexibility, the set of SCells affected by the implicit switching can also be RRC-configured.



Being able to bring SCell(s) in and out of dormancy relatively quickly is beneficial for power saving. Leveraging the BWP framework allows reusing many of the mechanisms (e.g. DCI signaling, BWP inactivity timer, etc), resulting in less specification work and impact.

Dormant BWP should be configured considering the following. There should be no PDCCH monitoring for the SCell when the dormant BWP is “active”. This can be achieved by CORESET / search space configuration for the BWP. CSI or SRS activities should have reduced duty cycle, and these can be configured accordingly for the BWP. RRM requirement can also be relaxed. Beam management can also be relaxed according to reduced CSI transmissions.

Proposal 5: SCell dormancy state should be supported in Rel-16 NR for power saving. The BWP framework should be enhanced to support operation with “dormant BWP”.

3.1.1.2.2 Semi-persistent search space configuration

In Section 3.4.1, a DCI-based search space set activation/deactivation scheme is proposed in the context of PDCCH monitoring reduction. In this scheme, the SS sets in the SCell remains inactive normally, and can be selectively activated by DCI signaling from the anchor SS set. When a SS set is inactive, the UE will skip PDCCH monitoring in the set.

In connection with the dormancy state, an SCell may be said to be dormant when all the configured SS sets within the cell are inactive. Hence, the DCI-based SS set activation/deactivation may be considered as an alternate signaling method for fast transition of SCell to the dormancy state.

Proposal 6: Semi-persistent search space configuration based on DCI-signaling can be considered to support fast transitioning in and out of SCell dormancy state.

3.1.1.3 Fast SCell activation / deactivation

The motivation for fast SCell activation/deactivation is so that the number of activated SCell can adapt more closely to the actual traffic load, thereby improving power efficiency. In NR, the timeline for SCell activation/deactivation can be optimized by using aperiodic TRS/CSI-RS and DCI-based signaling.

More details are discussed in our other contribution [16], where the following proposals are made:

Proposal 7: A-TRS/CSI triggering from SCell activation command should be supported in Rel-16 to greatly reduce SCell activation time.

Proposal 8: DCI-based signaling for SCell activation/deactivation should be supported in Rel-16, if other enhancements are also adopted to greatly reduce the overall activation timeline.

3.1.1.4 Latency Comparison

One difference with SCell dormancy state is that when a SCell is deactivated, UE does not have to perform any measurement or operations on the SCell, whereas for SCell dormancy state, UE may perform CQI measurements and reporting, albeit at a much sparser periodicity. It is expected that transitioning from dormancy state to activated state is still much shorter than transitioning from deactivated state, if it is implemented similar to BWP transition. Based on initial estimates, even for the best optimization scenario, SCell activation latency may be reduced to 7~12 milliseconds, but still quite large compared to SCell dormant BWP transition latency.

Scheme	Latency (15kHz SCS)	Comment
Dormant BWP switching	1~3 msec	BWP transition latency for multiple CC is to be decided in RAN4
Fast SCell activation	7~12 msec	Best optimization estimate
Rel-15 SCell activation (baseline)	24~84 msec	Upper range is for unknown cell. Lower range for known cell with SCell measurement cycle less than 160 msec.

One way the two proposed schemes can be differentiated is that SCell dormancy state targets faster (short-term) traffic variations, while fast SCell activation/deactivation targets slower (long-term) traffic variations.

3.1.2 Enhancement to BWP Adaptation

3.1.2.1 A-CSI-RS to assist BWP switching

Rel-15 BWP feature does not fully support triggering A-CSI request and DL BWP switch in the same slot. It is not explicitly disallowed from the spec, but functionally it does not work as intended. For the case an UL grant triggers A-CSI request as well as BWP change. The PUSCH transmission carrying the CSI report is indicated by k2 in the UL grant, but if this grant also triggers UL BWP switch, according to the spec, UE is not expected to receive or transmit for k2 slots. This means even if CSI-RS can be transmitted in DL BWP, UE is not expected to receive it.

During Rel-15 standardization, whether CSI-RS for non-active BWP should be measured for non-RRM purpose had been vigorously discussed and the conclusion is that CSI measurement is only performed in active BWP, with the exception for RRM for which gap-based measurement is allowed. Rel-16 BWP enhancement should not deviate from this basic decision.

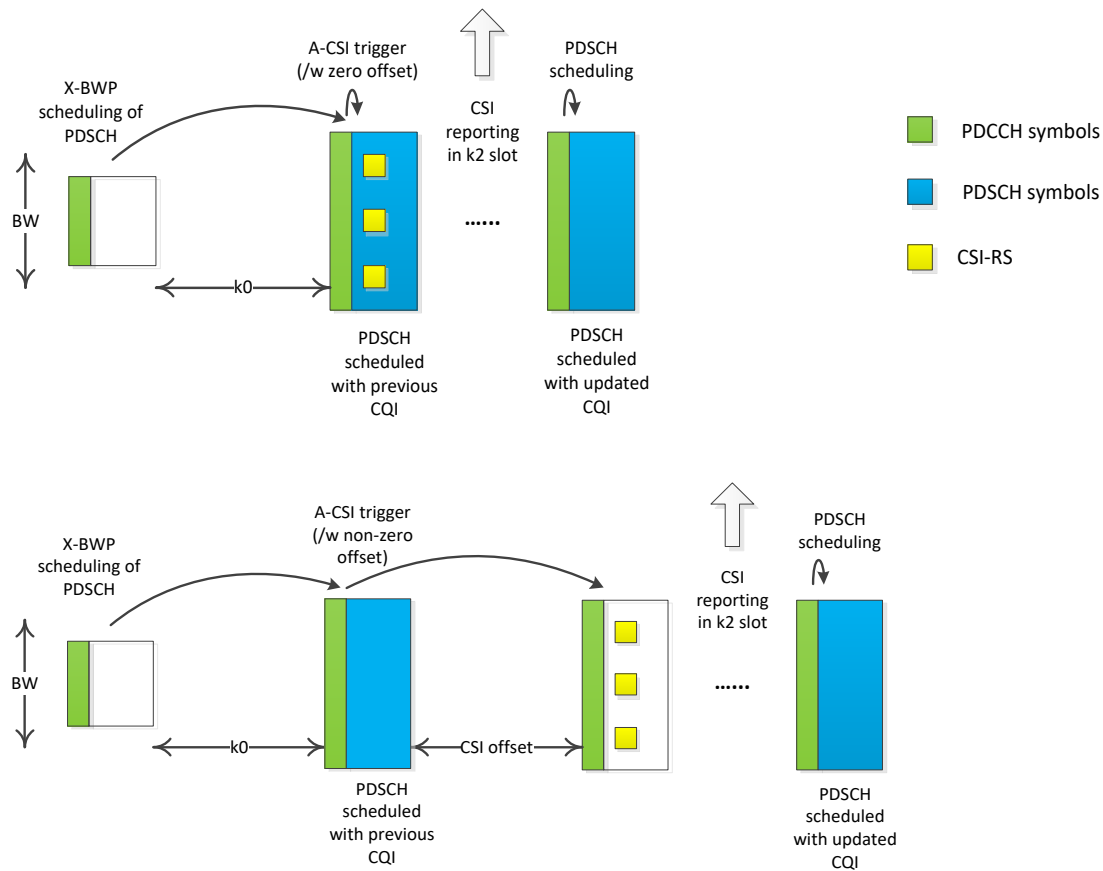
Observation 14: During Rel-15 standardization of BWP adaptation, there has been extensive and thorough discussion on whether CSI measurement in non-active BWP should be supported. In the end, the only such supported case is for RRM gap-based measurement. The decision was motivated by BWP adaptation for power saving use case and there is no need to reopen the discussion.

After a BWP switch, until updated CSI is measured and reported for the new BWP, the gNB still schedules with prior knowledge of the CQI, which can be based on the previous BWP (if the BWP are overlapping) or on the previous time when the new BWP was active. It helps performance for CSI to be measured and reported for the new BWP as soon as possible.

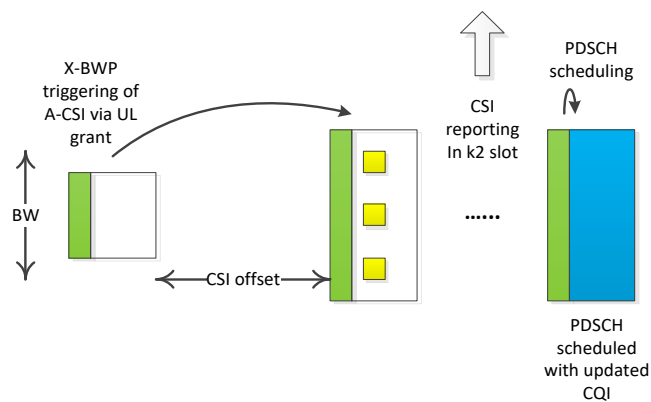
Observation 15: Rel-15 does not fully support triggering A-CSI request and BWP switch in the same slot. In some cases, A-CSI is delayed, and traffic performance may degrade, resulting in effectively higher BWP transition overhead.

If BWP transition overhead is too high, NW may reduce BWP adaptation usage, leading to lower power efficiency.

There exists inefficiency in how aperiodic CSI is handled in Rel-15 BWP framework. While it is possible to trigger A-CSI on the first slot in the new BWP, for Type-D QCL the A-CSI triggering offset would be non-zero and this may delay the earliest time CSI is updated for the new BWP. The following figures illustrate the timeline.



If the A-CSI trigger is issued at the same slot as the BWP switch trigger, CSI trigger offset can be concurrent with the BWP transition latency, and the overall timeline can be shortened, as illustrated below.



Observation 16: It is beneficial to have a L1 signal that can trigger BWP switch and at the same time trigger A-CSI measurement and reporting. Given that DCI signaling is already supported for BWP switching in Rel-15, that can be used as the starting point.

For above operation to be fully supported and functional, several issues need to be addressed, including:

- Cross-BWP triggering of A-CSI measurement and reporting should be fully supported in the spec

- How to define the relationship between A-CSI triggering offset and BWP transition latency
 - A-CSI triggering offset should be configured to be larger than BWP transition latency requirement
 - If A-CSI triggering offset can be indicated directly in DCI, it can be expected to accommodate BWP transition latency.
- A-CSI-RS configured for the new DL BWP is triggered by the DCI field sized for the previous DL BWP.
- For paired spectrum, UL grant that triggers A-CSI does not also implicitly trigger DL BWP switch. Ways to trigger DL BWP switch in the same slot as the UL grant should be considered.

Proposal 9: Cross-BWP triggering of A-CSI measurement and reporting should be supported in Rel-16. Triggering of aperiodic CSI and BWP switch in the same slot should be fully functional.

3.1.2.2 Association of BWP and DRX operation

BWP adaptation in Rel-15 operates fairly independently from DRX operation. For example, going into DRX state does not alter the active BWP. On the other hand, given that both features are intended for power saving, it would be more optimized to couple the operation of the two features in some way. DRX can be viewed as a power saving technique in the time domain by gating PDCCH monitoring and other Rx operations between ON and OFF states, while BWP can be viewed as a technique that gives control for different “levels” of power saving for the DRX ON states.

The following BWP and DRX state association may be beneficial.

- One of the configured BWP can be designated as the “starting BWP” for DRX ON duration. If there is a power saving BWP (i.e. with narrow BW and large k_0), UE may want to wake-up on that BWP and check for grants. If there is no grant, UE returns to DRX while having spent less power for the ON duration. If this BWP-DRX association is not in place, UE would start on whichever BWP that was previously active from the previous DRX cycle, and the power saving is not guaranteed.
- Above can be extended if there is a wake-up signal prior to the DRX ON duration. In this scenario, it makes sense to associate a high performance (i.e. with wide bandwidth) BWP to the ON duration. This is because the wake-up signal would tell UE to skip the ON duration if there is no scheduling, so if UE is to be active during the ON duration, it means it is expecting scheduling and use of wide BW would be more efficient.
- It may also be beneficial to tie certain BWP inactivity timer behavior to DRX state. For example, resetting it during DRX state so that the timer’s behavior is more predictable at the beginning of the next cycle. Also, if ON duration is typically configured to be short, BWP inactivity timer may be configured to start counting only when DRX inactivity timer is started.

Proposal 10: One of the configured BWPs can be designated as the BWP which is always activated when UE starts for DRX ON duration. Other association between BWP and DRX operation can be considered for Rel-16.

3.1.3 Cell Grouping for Power Saving

In our companion paper submitted to RAN2 [19], we explain why each MAC entity having only a single DRX configuration in all scenarios can become inefficient in Rel-16, when UE can be configured with a set of serving cells operating in different frequency bands.

For example, FR1 cells are likely used for scheduling and applications that require extended coverage (e.g. VoNR). Those types of applications typically do not have large bursts of transmission but require low latency. This implies that they could benefit from a DRX configuration with short DRX cycles to meet the delay required, and short on duration and DRX inactivity timer to maximize power saving. On the other hand, FR2 cells are more likely used for high data-rate, very bursty applications, which often have random and long period of inactivity between data bursts. To maximize power saving, it is more efficient to have a DRX configuration that has long DRX cycle. And their ON duration and DRX inactivity timer can be set longer than those for FR1 cells to better handle randomness in the duration of data burst.

In addition, in some UE implementations, FR1 and FR2 serving cells may be supported by different transceivers whose power are managed independently. In that scenario, it is desirable for each serving cell supported by different transceivers to have different DRX configurations. On the other hand, microsleep for wideband RF transceiver covering a group of cells is only feasible when all the cells within the group are not scheduled. gNB can try to align unscheduled slots across cells within a group as much as possible, in order to maximize microsleep opportunities.

Observation 17: A group of cells sharing the same wideband RF transceiver may be able to exercise more microsleep if gNB scheduler can align unscheduled slots across cells.

Proposal 11: Serving cells with different uses can be assigned to different groups, and different cell groups can be configured with different DRX parameters. gNB can time-align scheduling across cells within a group to maximize microsleep. UE may recommend grouping of cells via UE assistance information feedback.

3.2 Adaptation to the variation in time

3.2.1 Cross-slot Scheduling

Adapting instantaneous power consumption to per-slot scheduling requires the use of a technique called “microsleep”. If the UE knows a-priori the range of symbols do not carry transmissions that the UE may be expected to receive or transmit, the UE can put its RF and portion of the front-end hardware into power saving mode for those symbols. With cross-slot DL scheduling, if it is known from previous slot’s DL control that the current slot is not scheduled, microsleep can start as soon as the last DL control symbol ends. The portion of microsleep can be significantly extended compared to the same-slot scheduling baseline, for which microsleep cannot start until PDCCH blind decoding completes without decoding any grant.

One subtlety to point out is that it is not enough for network to be able to schedule with $k_0 > 0$ by DCI indication of k_0 . If $k_0 = 0$ is among the semi-statically configured k_0 candidates in pdsch-TimeDomainAllocationList table, UE still cannot implement extended microsleep. This is because until the UE can finish blind decoding of all PDCCH candidates, it cannot be sure that there would be no DL assignment indicating $k_0 = 0$. To enable extended microsleep, what matters is that none of the k_0 candidate values semi-statically configured in pdsch-TimeDomainAllocationList table should correspond to $k_0 = 0$.

$K_0 > 0$ configuration (i.e. cross-slot scheduling) is beneficial for UE power saving but it comes at slight expense of latency. It would be desirable to be able to switch to same-slot scheduling ($k_0 = 0$) mode during a traffic burst. However, reconfiguration of pdsch-TimeDomainAllocationList has to be done through RRC signaling which is slow and takes a lot of overhead.

The same discussion generally also applies to UL scheduling (i.e. PUSCH scheduling offset indicated by k_2). To guarantee cross-slot scheduling for UL, pusch-TimeDomainAllocationList table should have all entries with $k_2 > 0$. For most cases, UE capability which defines N_2 would dictate how fast UL grant processing needs to finish before the start of PUSCH symbol. But if $k_2 > 0$ can be ensured by configuration, it would certainly further relax grant processing timeline and would likely lead to power saving benefit.

3.2.1.1 Rel-15 Support for cross-slot scheduling

3.2.1.1.1 Issue with Rel-15 A-CSI triggering offset

In Rel-15, cross-slot DL scheduling for increased microsleep is supported by configuring pdsch-TimeDomainAllocationList table without any entries with $k_0 = 0$. However, current NR specification contains a “bug” that may prevent the support. In order to put RF and frontend to sleep, not only PDSCH scheduling should be considered, but A-CSI request and CSI-RS should also be considered. If CSI-RS can come in the same slot as the UL grant that triggers the A-CSI request, this means UE cannot determine to start microsleep until all PDCCH blind decoding is done. Similar to $k_0 > 0$ condition for PDSCH, for A-CSI request, the *triggering offset*, which defines the slot delay for the CSI-RS relative to the PDCCH, should be greater than zero.

However, in RAN1#92, the following was agreed:

For CSI acquisition, aperiodic CSI-RS triggering offset can be 0, 1, 2, 3, 4 slots.

- *If all the associated trigger states do not contain QCL Type D information, aperiodic CSI-RS triggering offset is fixed to zero*

This means, for certain QCL Type configuration, the A-CSI-RS trigger offset is zero and generally, extended microsleep for cross-slot scheduling would not be feasible. The issue should be corrected in the spec.

Proposal 12: For CSI acquisition, allow non-zero aperiodic CSI-RS triggering offset in the case that all the associated trigger states do not contain QCL Type D information.

3.2.1.1.2 BWP-specific configuration of TDRA table

Fast adaptation between cross-slot scheduling and same-slot scheduling can be done with BWP adaptation. Because the TDRA table (i.e. pdsch-TimeDomainAllocationList) is BWP-specific, some BWP can be configured with all entries with $k_0 > 0$, whereas some BWP can be configured with entries containing $k_0 = 0$. Then, switching between BWP can achieve the result of adapting between minimum $k_0 > 0$ and minimum $k_0 = 0$.

During BWP switch, there is a subtlety that for the DCI that triggers the BWP switch, the TDRA table configured for the target BWP is used. Because a UE does not know a-priori when it would receive a BWP-switching DCI, it may not be generally feasible to guarantee that schedulable k_0 is always greater than zero, even if minimum k_0 is configured to be greater than zero for the current BWP. There are two approaches to work-around this problem:

- (i) For a BWP other than the current BWP (intended to support cross-slot scheduling), if $k_0 = 0$ entry should be configured in the corresponding TDRA table, assign the entry with a higher index such that it is not addressable by the bit-width of the frequency domain RA field of the current BWP.
 - For example, frequency domain RA field has bitwidth of 1 bit for the current BWP. For the other BWP, make sure the $k_0 = 0$ entry (if any) should be assigned index 2 or larger.
- (ii) BWP transition time defined in the spec should be large enough such that the $k_0 = 0$ entries in other BWPs are considered non-schedulable from the current BWP. This is because there is a requirement that the indicated k_0 for cross-BWP scheduling has to accommodate BWP transition time (which is defined to be slightly more than 2 msec by RAN4 for Type 2 switching; Less for Type 1).

Above discussion generally applies to UL as well with minimum k_2 and pusch-TimeDomainAllocationList as the counterparts of minimum k_0 and pdsch-TimeDomainAllocationList.

Given that crossing one-slot for scheduling typically is sufficient to get extended microsleep benefit, and BWP transition time is at least one slot, only (ii) is needed. (i) can be used for guaranteeing $k_0 > x$ where x is very large. For subsequent discussion, Approach (ii) is assumed but the same discussion can always be extended for (i) as well.

3.2.1.1.3 Work-around for default and common TDRA tables

For DCI in UE-specific search space, the dedicated TDRA table is used if provided in pdsch-Config. To enable cross-slot scheduling for power saving, care can be taken to ensure all the entries are with $k_0 > 0$. For DCI scrambled with SI/P/RA/TC-RNTI, monitored in Type 0/0A/1/2 CSS, the default or common TDRA table is used. Also, for DCI scrambled with CS/MCS-C/C-RNTI in CSS in CORESET 0, the common TDRA table is used. The default TDRA table is fixed in the spec [Section 5.1.2.1.1 in TS 38.214] and contains $k_0 = 0$ entries. The common TDRA table can be configured and provided in pdsch-ConfigCommon. When any one of these kinds of DCI is being monitored in a slot, it may not be possible to guarantee $k_0 > 0$ in the TDRA table. Power saving by extending microsleep may not be feasible.

To work-around this problem, if only the common table is used in addition to the dedicated table, it can also be configured with $k_0 > 0$ for all entries. If this is not feasible, or if the default table is in used, at least the problem should be localized in time, because the default and/or common tables may not be “active” majority of the time. The TDRA table is only “active” when a DCI with corresponding RNTI and search space type is being detected as a PDCCH candidate. But the duty cycle for monitoring DCI scrambled with SI/P/RA/TC-RNTI should be small, so even if $k_0 = 0$ has to be supported it would not impair $k_0 > 0$ power saving for the majority case. There are also ways to avoid or reduce monitoring DCI scrambled with C-RNTI in CORESET 0, in case $k_0 > 0$ cannot be guaranteed in the common TDRA table.

3.2.1.2 Minimum scheduling offsets for Rel-16

3.2.1.2.1 Motivation

Management of different minimum k_0 across BWP is not very straight-forwarded in Rel-15. This is due to the aspect that for cross-BWP scheduling (which triggers BWP switch), the target BWP's TDRA table would be used. Minimum k_0 is a function of all the entries in the TDRA table. Also, UE does not know a-priori when and which BWP it would be triggered to switch to (in case there are more than two BWPs configured). To simplify the discussion, let's consider the case only the

dedicated TDRA table provided in pdsch-Config is used (See discussion in Section 3.2.1.1.3 for the other case). The minimum k_0 that a UE has to be prepared to handle can be expressed as the following:

$$\min(\min_k0(\text{BWP}_0), \max(\text{BWP transition time in slots}, \min_{n=1..numBWP-1}(\min_k0(\text{BWP}_n))))$$

where

- $\min_k0(\text{BWP}_x)$ is the minimum k_0 across all entries in the TDRA table for BWP_x , where $x=\{0, \dots, numBWP-1\}$
- $numBWP$ is the number of configured DL BWP
- Without loss of generality, BWP_0 is assumed to be the current BWP

Above can be generalized to minimum k_2 as well. Overall, minimum k_0 and minimum k_2 can be discussed more generally as “minimum DL scheduling offset” and “minimum UL scheduling offset” to cover requirements for aperiodic CSI triggering, etc.

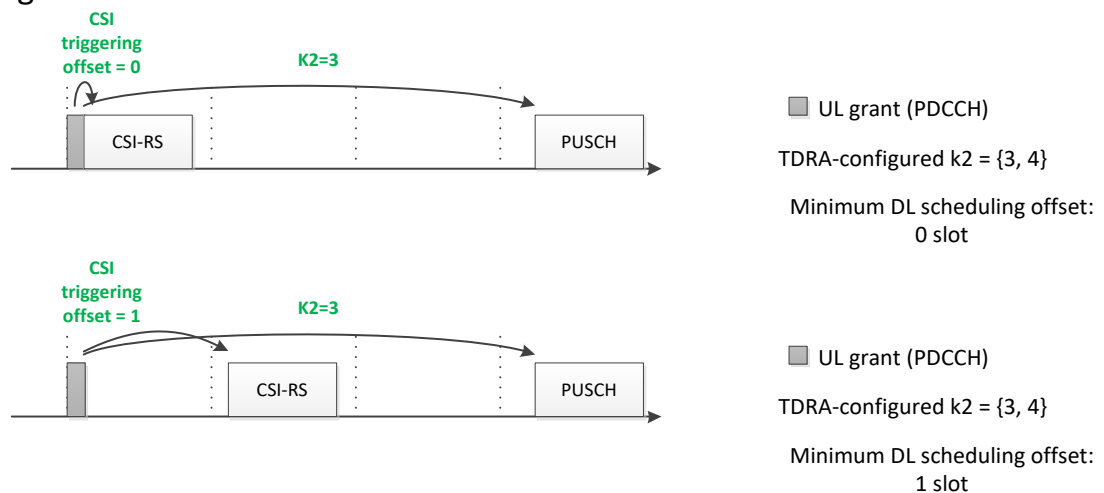
For Rel-16, minimum scheduling offset configuration should be simplified. It would be much more straight-forward to have explicit configuration of the minimum DL scheduling offset. This minimum DL scheduling offset serves the following usage:

- It explicitly controls the minimum k_0 that UE is expected to handle for PDSCH scheduling, even for cross-BWP scheduling (i.e. triggering BWP switch)
- It defines the minimum timing offset for aperiodic CSI-RS triggering
- The only exception is for PDSCH scheduled with the default TDRA table
 - E.g. SIB, for which the k_0 can be smaller than the minimum DL scheduling offset

Similarly, a minimum UL scheduling offset can be explicitly configured, serving UL scheduling usage.

For aperiodic CSI triggering, the actual slot timing of the CSI-RS relative to the triggering grant can be a function of the minimum DL scheduling offset and the configured A-CSI triggering offset, as illustrated below.

A-CSI triggering:



Proposal 13: NR should consider explicit configuration/signaling of minimum DL scheduling offset and minimum UL scheduling offset for Rel-16.

Proposal 14: Timing determination for A-CSI-RS relative to the triggering grant can be a function of the configured A-CSI triggering offset and the minimum DL scheduling offset.

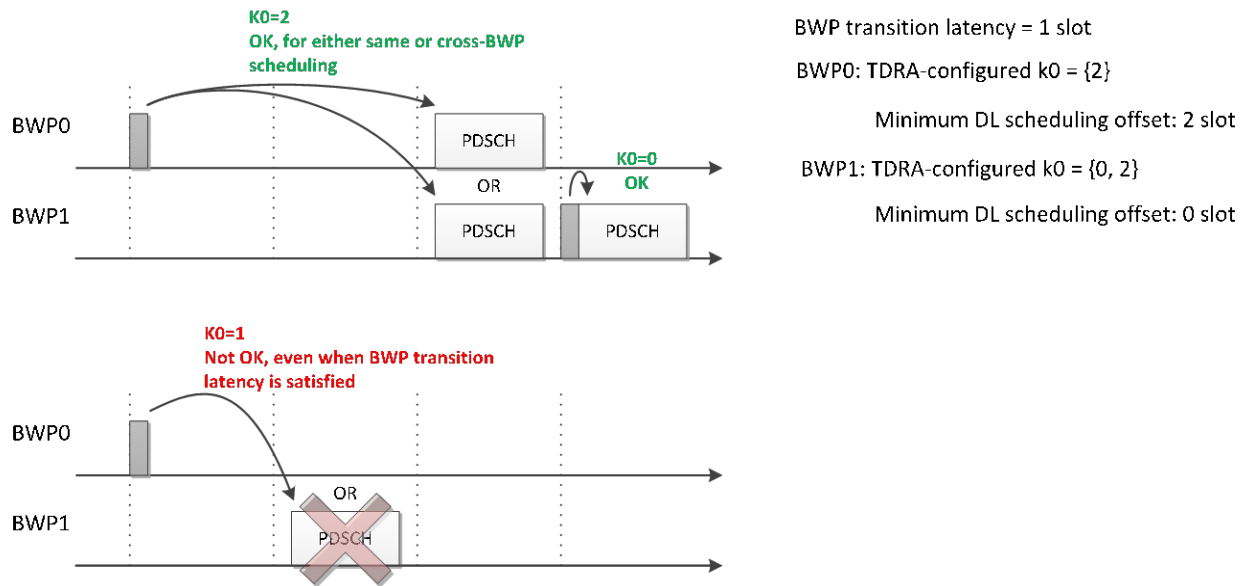
Proposal 15: A-CSI triggering offset can be indicated in the DCI with the A-CSI request, and satisfy the minimum DL scheduling offset constraint.

3.2.1.2.2 With BWP adaptation

Minimum DL/UL scheduling offsets should be attributes of BWP configurations. In addition to TDRA table configuration per DL BWP, the minimum DL scheduling offset can be configured per DL BWP; The same can be done for minimum UL scheduling offset and UL BWP. The minimum scheduling offset for the currently active BWP is applied to same-BWP scheduling DCI and potentially any cross-BWP scheduling DCI (which triggers a BWP switch). When there is a BWP switch, the new BWP's minimum scheduling offset starts to apply when the BWP becomes active. This facilitates fast adaptation between cross-slot scheduling and same-slot scheduling to achieve good tradeoff between power saving and traffic performance.

Support for multiple minimum DL scheduling offsets per DL BWP or multiple minimum UL scheduling offsets per UL BWP can be further considered as an enhancement.

Multiple BWP:



Proposal 16: Minimum DL scheduling offset or minimum UL scheduling offset configuration is per DL or UL BWP respectively, and RRC-configured as part of the BWP configuration.

UE may be configured with “power saving BWP” intended for starting ON duration. The power saving BWP may be configured with narrow BW and large minimum scheduling offsets for DL and UL. If there is no traffic for the UE, after ON duration finishes UE returns to DRX after spending little power. If there is traffic for the UE, it can be issued a cross-BWP scheduling DCI to switch to another BWP with wide BW optimized for data transfer. The large minimum DL/UL scheduling offset condition is necessary because this ensures that even if a grant is detected, UE is not expected to receive the PDSCH or transmit the PUSCH until after some time guaranteed by the offsets. This way, UE does not have to warm-up the hardware for PDSCH or PUSCH processing until a grant is detected. This helps power saving – the power level is estimated to be even lower than simply extending microsleep with cross-slot scheduling. Large values for both minimum DL/UL scheduling offsets have to be used, the smaller one among the two would be the earliest time UE has to warm-up hardware in addition to those necessary for PDCCH processing and decoding. It would make sense to configure minimum scheduling offset with similar values for DL and UL.

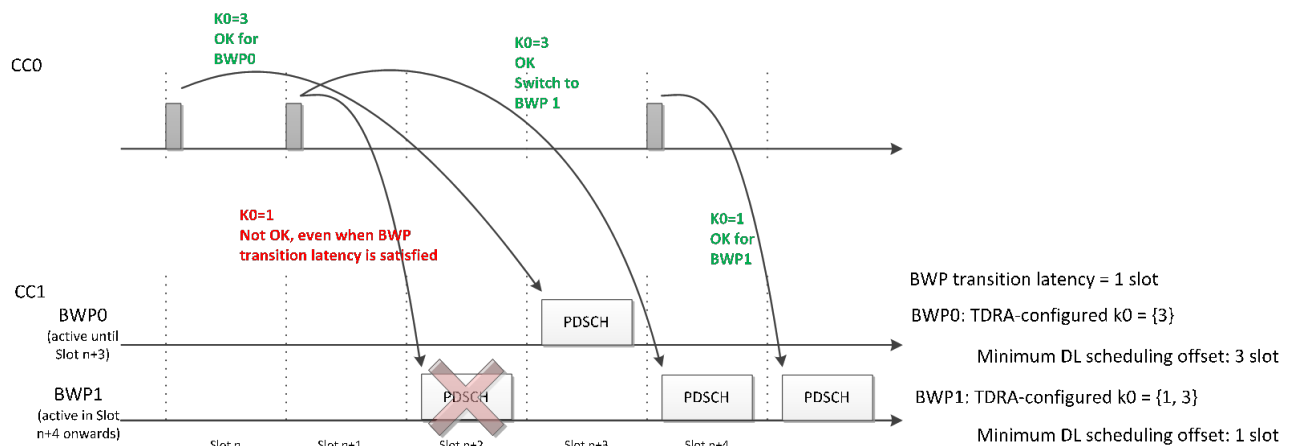
Observation 18: For the power saving BWP used for DRX wake-up, large minimum DL and UL scheduling offset facilitates additional power saving. Minimum DL and UL scheduling offset should be configured to be similar in value.

3.2.1.2.3 With cross-carrier scheduling

According to current NR specification, search space configuration for cross-carrier scheduling is based on the currently active BWP on the scheduled carrier. There is a linkage rule for the search space defined for the scheduled carrier to that for the scheduling carrier. Similar to self-scheduling, the minimum k_0 is determined based on the configured TDRA tables across all schedulable BWP on the scheduled carrier, along with any additional conditions such as BWP transition latency.

For the cross-carrier scheduling feature, there is an even stronger motivation for introducing explicit minimum scheduling offsets. If traffic on SCell is light, there could be long gaps of no scheduling on the SCell. UE may save more power by operating in lower power mode (e.g. at reduced clock / voltage for the baseband); It may even choose to suspend processing related to SCell while it is not being scheduled. However, for such power saving to be feasible, it would be a prerequisite to guarantee a relatively large scheduling delay from the scheduling carrier (e.g. PCell) to the scheduled carrier (e.g. SCell), such that there can be enough time for hardware to transition to higher power mode to process the scheduled operations on the SCell. Similar to self-scheduling, for Rel-15, to some degree, this can be achieved by careful configuration of minimum k_0 across the BWP of the scheduled carrier. Explicit minimum scheduling offsets applied to cross-carrier scheduling would make the feature much more usable.

Cross-Carrier Scheduling



In above example, BWP0 of CC1 can be the “power saving BWP” as it is configured with a large minimum DL scheduling offset. It can be used most of the time when traffic is sparse. When there is more traffic, BWP1 of CC1 can become the active BWP and a smaller minimum DL scheduling offset can be used for lower latency.

On a related note, minimum k_0 offset for cross-carrier scheduling with different numerology is a promising proposal to resolve some of the issues in buffering requirement and causality processing requirement [17]. Explicit minimum scheduling offset scheme would be a unified mechanism that works for both self-carrier scheduling and cross-carrier scheduling.

Minimum scheduling offset determination should be based on UE feedback, because the amount of scheduling delay that facilitates power saving is UE-implementation dependent.

Proposal 17: Minimum scheduling offset determination should be based on UE capability signaling and/or UE-assistance framework.

3.2.2 Same-slot Scheduling

3.2.2.1 TDRA configuration restriction

It has been suggested that TDRA configuration can facilitate microsleep for same-slot scheduling, for example, by applying a TDRA configuration restriction to ensure a gap between PDCCH reception and PDSCH transmission known to the UE.

In this section, several clarifications on above technique will be made and important distinction with cross-slot scheduling (facilitated by minimum scheduling offset) will be made.

First, it should be pointed out that when there is a certain amount of data to be transmitted, it is generally more energy efficient to minimize the time span of the transmission. This is due to fixed power overhead even when UE is not transmitting / receiving if short gaps are introduced, for example, the power level during microsleep is 45 power units according to the agreed power model. If UE can finish transmitting / receiving data as quickly as possible, there would be more likelihood for entering and spending more time in deeper sleep state (e.g. deep sleep at 1 power unit). Introducing gaps in the transmission tends to increase overall energy consumption for the transmission, thereby reducing overall power efficiency.

Therefore, the suggestion to ensure a gap between PDCCH reception and PDSCH transmission for power saving actually presumes a niche use case, where

1. The amount of data to be transmitted is very small (i.e. transmission for a fraction of a slot would be sufficient)
2. The data also needs to be low latency such that same-slot scheduling has to be used

To elaborate, it should be obvious that if the amount of data is not small, but if a gap needs to be ensured between PDCCH and PDSCH, the overall time span of the transmission may have to last over many slots. This would introduce additional energy consumption and latency; A better approach is to configure full slot transmission and use cross-slot scheduling to maximize microsleep if the traffic is bursty. It should be pointed out that the additional latency for introducing gaps is multiplicative of the amount of data to send (because it has the same effect as throughput reduction), whereas the additional latency due to cross-slot scheduling is fixed and additive (i.e. just the scheduling delay).

Observation 19: Ensuring a gap between PDCCH reception and PDSCH transmission in TDRA table configuration for same slot scheduling may have some drawbacks compared to the baseline of full-slot, back-to-back transmissions

- Resulting in lowered peak throughput, and generally, hurts energy efficiency for typical traffic.
- Adds latency which would be multiplicative of the amount of data to send

Obviously, for the case if conditions (1) and (2) are both met, and PDSCH over partial slot is sufficient for the data, then it would be the most power efficient to place the gap between PDCCH and PDSCH. But more importantly, this is a niche usage scenario and should not be discussed as an alternative to cross-slot scheduling.

Cross-slot scheduling can save power without any additional signalling from the network. It can work efficiently when traffic activity is high or low. When traffic activity is low and many slots are unscheduled, microsleep can be maximized. When traffic activity is high, back-to-back full slot transmission can be scheduled. With same-slot scheduling approach, when traffic activity is low, TDRA should be configured to guarantee a gap between PDCCH and PDSCH within slot. When traffic activity is high, TDRA configuration must be adjusted to remove the gaps and support full-slot transmission.

Observation 20: Compared to same-slot scheduling with TDRA table restriction, cross-slot scheduling for UE power saving has wider applicability and does not rely on any additional signalling from the network.

Another aspect is UE implementation. Generally, PDCCH processing timeline is highly UE implementation dependent. If a gap should be introduced between PDCCH and PDSCH, it is unclear how much would be sufficient. On the other hand, cross-slot scheduling requires PDCCH processing completion within slot-granularity instead sub-slot-granularity, and this should give more flexibility to UE implementation.

The latency to enter and exit microsleep and the extent of microsleep is also highly UE implementation dependent. For same-slot scheduling, the gap potentially available for microsleep is only a fraction of a slot when the slot is scheduled. In reality, there is non-negligible overhead and latency in entering and exiting microsleep, reducing the actual microsleep gain during short gaps within a slot.

Observation 21: The amount of UE power saving achievable based on same-slot scheduling with TDRA table restriction is highly dependent on UE implementation, whereas for cross-slot scheduling, the dependency is more relaxed leading to more robust gain.

3.2.2.2 TDRA Entry Selection to Facilitate Cross-Slot Scheduling

It has been suggested that dynamic or semi-static selection of TDRA entries can be the mechanism to limit scheduling options known to the UE so that microsleep durations can be maximized. Because the configured values of k_0 is part of the TDRA table, it has also been further suggested that this scheme can be used to facilitate cross-slot scheduling for power saving. While this looks convincing on the surface, careful examination of operation involving BWP operation reveals several issues.

TDRA tables are BWP specific; If there should be selection of TDRA entries for enabling/disabling, the selection would have to be BWP specific. The most sensible schemes would be to provide a list of the indices of the disabled entries, a.k.a. “disable-list”, or a bitmap corresponding to all the entries in the table and the bit position is associated with the index of the entry, and each bit indicates disable/enable status of the associated entry.

First, supporting dynamic/semi-static selection of TDRA entries does not address the A-CSI triggering offset issue, because TDRA is not used to determine CSI-RS timing. On the other hand, supporting an explicit minimum DL scheduling offset parameter opens up some ways to address that issue as discussed in Section 3.2.1.2.

Observation 22: TDRA entry selection does not resolve the A-CSI triggering offset issue.

Second, TDRA entry selection approach has some complications and inefficiency in supporting BWP operation (and applies equally to cross-carrier scheduling). Suppose the objective is to have minimum k_0 not smaller than x for the current BWP, denoted as BWP0. Based on the approach of selection of TDRA entries, the “disable-list” should contain the indices of the entries with $k_0 < x$. However, suppose another BWP, denoted as BWP1, is configured with some TDRA entries smaller than x . Because it is possible that UE receives a cross-BWP scheduling DCI that triggers switch from BWP0 to BWP1, the TDRA table for BWP1 could come into play even when the active BWP is BWP0. To guarantee the scheduled k_0 would be at least x , one way is to also disable the entries with $k_0 < x$ for BWP1 by the “disable-list” for BWP1. Extending further, there can be another configured BWP, or even more. The same has to be done to each of the configured BWP’s “disable-list” just because $k_0 < x$ should be disallowed when active BWP is BWP0. Suppose a switch to high power BWP occurs and $k_0 \geq x$ condition is no longer needed. Extra signalling overhead would be needed to update each and every BWP’s “disable-list”. On the other hand, the scheme described in Section 3.2.1.2 requires no extra signalling overhead on top of the existing BWP switch signalling to support the same objective and operation.

Observation 23: To guarantee $k_0 \geq x$ using TDRA entry selection with multiple configured BWP, update to TDRA entry selection for the non-active BWPs may be needed, leading to extra signalling overhead.

3.2.3 Multi-slot Scheduling

Some energy can be saved if a “back-to-back” PDSCH transmission over a number of slots can be all scheduled at the beginning by multi-slot scheduling. Based on the agreed power model, the saving for a PDSCH-only slot is about 6% to 7%. It is not substantial and given other potential complications with multi-slot scheduling, our view is that the priority for this power saving scheme is not high.

For cross-carrier scheduling from small to large SCS, multi-slot scheduling would be a natural by-product and it is important even from power efficiency perspective that back-to-back scheduling for the scheduled carrier is possible.

On the other hand, we think increasing PDCCH monitoring periodicity can provide more gain to UE power saving, because sleep time can be extended. Combining the two techniques may be considered.

3.3 Adaptation to DRX operation

3.3.1 Preparation period for C-DRX

In NR C-DRX operation, in addition to the regular wake-up for ON-durations, additional wake-up may be required in the middle of C-DRX off periods. For example, the UE may rely on SS/PBCH blocks for time/frequency synchronization and beam management (for FR2) during C-DRX operation. However, since SSBs are sparse in time and not necessarily aligned with ON-durations, the UE may need to wake up during off periods to measure the SSBs. Obviously, any additional wake-

up during off periods imposes overheads for ramping-up RF front-ends and prevents the UE from going into deeper sleep, which can result in significant power penalty.

A potential solution to reduce the number of additional wake-up may be utilizing CSI-RS. In Rel-15, CSI-RS can be transmitted during C-DRX ON-durations. For fast warm-up of UE, it is preferred to transmit the CSI-RS in the beginning of ON-durations. However, measuring CSI-RS and performing required processes (e.g., tracking, beam management, or CSI acquisition) involves their own latency, and the UE's link performance may be impacted until those processes are finished, especially when the ON-duration is very short.

To overcome the issue, in Rel-16, a pre-wake-up (PWU) window can be considered at a little ahead of time before ON-durations, as illustrated in Figure 9. For further flexibility, the PWU window may only be present in a subset of C-DRX cycles. The primary use of the PWU window is carrying CSI-RS resources to facilitate UE's timely warm-up, but there may be other versatile applications of the PWU window that can offer gains both in power consumption and link performance.

Proposal 18: Pre-wake-up window is considered as a C-DRX enhancement scheme for power saving and performance improvement.

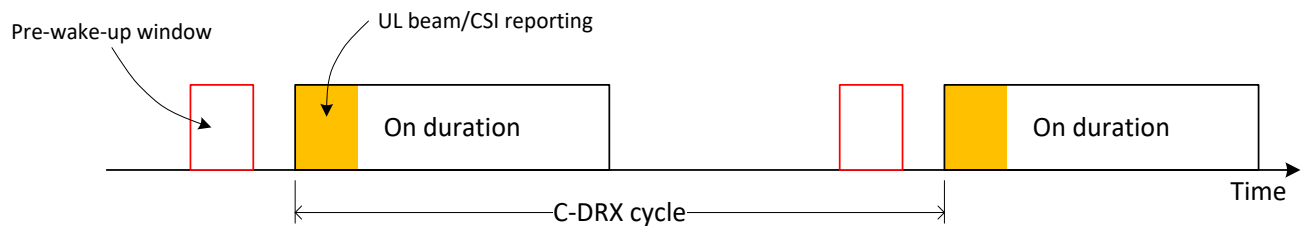


Figure 9: C-DRX timeline with pre-wake-up window

3.3.1.1 PWU windows for CSI-RS transmission

In NR, there is no always-on reference signal, such as CRS in LTE, which the UE can use to warm-up the model hardware during C-DRX operation. Thus, the PWU windows can be utilized for UE (or group of UEs)-specific CSI-RS transmission from gNB. The UE can measure the CSI-RS resources and adjust its status to get ready for the upcoming ON duration. Depending on operating scenarios (e.g., frequency range, C-DRX cycle, UE mobility, etc.), the CSI-RS transmission may include one or more of different types, such as CSI-RS for tracking, CSI-RS for beam management, and/or CSI-RS for CSI acquisition.

For example, if the UE is configured to receive a TRS during the PWU window, it can recover AGC, time and frequency synchronization before the ON duration starts. Similarly, if the UE measures CSI-RS for BM in the PWU window, it can adjust Rx beams and, if required, perform fast beam reporting when the ON duration starts. Lastly, if the UE receives CSI-RS for CSI acquisition, it can measure and process CSI during the PWU window and report it during the ON duration.

Compared to the conventional schemes, which only relies on SSBs during C-DRX off periods or CSI-RS during C-DRX ON durations, the PWU window-based scheme can improve the power consumption as well as the latency. As analyzed in [15], the power saving gain may mostly come from the reduced number of additional wake-up during C-DRX off periods, and the latency gain may be from the shortened turnaround time by allowing fast beam/CSI reporting in the beginning of the ON duration. As a result, the benefit of the PWU window may be more substantial for the longer C-DRX cycle and the shorter ON duration.

3.3.1.2 PWU Window as WUS Occasion

In [9], it was discussed that the transmission of a wake-up signal (WUS) in a WUS occasion can lead to significant power savings due to the UE avoiding unnecessary wake-ups during the CDRX ON duration. For example, the gNB may choose to transmit a WUS within a WUS occasion only when it has data to transmit to the UE, subsequently, by detecting the presence of the WUS within the WUS occasion, the UE can decide whether to skip the next ON duration or not. Also, compared with control and data reception during the DRX-ON duration, the reception of the WUS in the PWU window could be optimized by using fewer number of Rx antenna, reduced bandwidth (e.g., power saving BWP) or more relaxed

timeline [9]. Therefore, due to the optimized reception of WUS and skipping PDCCH monitoring when there is no upcoming PDSCH transmissions, the UE is able to realize significant power savings compared to scenarios without WUS.

The key principle of WUS lies in the two-stage wake-up, which required a preparation time, a “WUS offset”, between the 1st and 2nd stages of wake-up. In that sense, the PWU window can be used as a WUS occasion for UEs employing WUS schemes.

This WUS could be based on L1 signaling such as CSI-RS and PDCCH. In [9], our views on WUS proposal are presented with extensive information regarding power saving gains, timeline, implementation options and specification impact.

Observation 24: CSI-RS and WUS Transmissions in the PWU has the potential to provide significant UE power savings.

3.3.1.3 CSI-RS Transmission and WUS Occasion in PWU Window

Separate Transmissions

Given the aforementioned benefits of CSI-RS transmission and WUS in the PWU window, these signals can be transmitted separately within the PWU. Separate transmission offers more scheduling flexibility, especially, when the periodicities of the signals are different. For example, in some scenarios, the CSI-RS for tracking or beam management may be transmitted in a PWU window of every y (>1) C-DRX cycles, while the WUS can be transmitted potentially in every PWU window as long as there is upcoming data transmission for the DRX ON duration.

In Figure 10, we illustrate an example where TRS and a PDCCH-based WUS are transmitted in the same PWU window. The TRS may come before, after, or even frequency multiplexed on the same time resource as the PDCCH-WUS. However, placing the TRS earlier than the PDCCH-WUS, the UE can recover synchronization first, which is beneficial both for the PDCCH-WUS detection and potentially for the next ON duration, in case the UE needs to wake up.

Once the UE receives the PDCCH-WUS, it prepares the receive chain for PDCCH reception in the upcoming DRX ON duration. This preparation may include BWP switching from a power saving BWP, which is configured for the PWU window, to another BWP. In the case that the gNB requires up-to-date CQI information before scheduling data, the gNB may trigger A-CSI transmission to the UE at the beginning of the DRX ON duration. Some details of the A-CSI-RS measurement steps (as shown in Figure 10) and some enhancement has been discussed in Section 3.1.2.1.

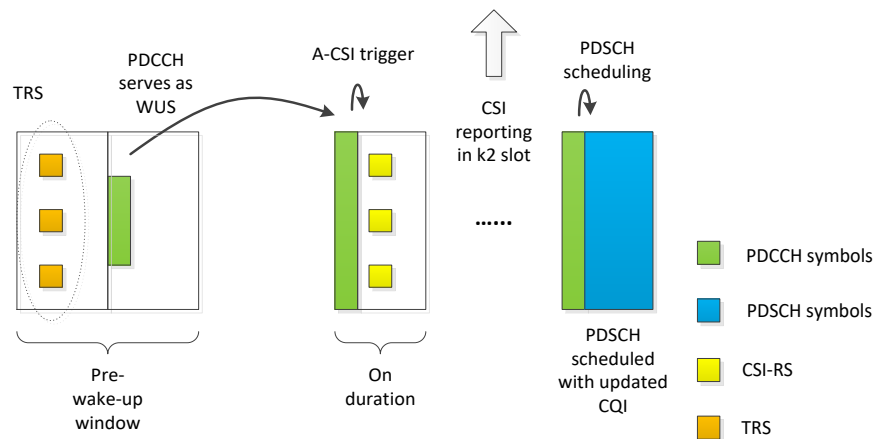


Figure 10: TRS and PDCCH Signaling in PWU Window

Single Transmission

For more efficient use of resources, instead of transmitting separate CSI-RS and WUS in the PWU window, a single signal may be considered to perform both functionalities. In other words, as proposed in our companion paper [9], the CSI-RS waveform can be used as a WUS to indicate upcoming data transmission, and the UE can also use the CSI-RS for measurements, tracking, or beam management purposes. An example of this scheme is illustrated in Figure 11. In this same figure, we also show that, in case the gNB desires up-to-date CQI, it could trigger A-CSI transmission to the UE.

It is noteworthy that, while the single transmission for CSI-RS and WUS is resource-efficient, it is rather restrictive in some cases. For instance, although a CSI-RS-like waveform is desirable to perform both functionalities, it is less robust than PDCCH-WUS, in general. Also, PDCCH-WUS transmitted from a power saving BWP may deliver further information, such as BWP switching along with cross-BWP A-CSI triggering, which is not possible with CSI-RS-WUS.

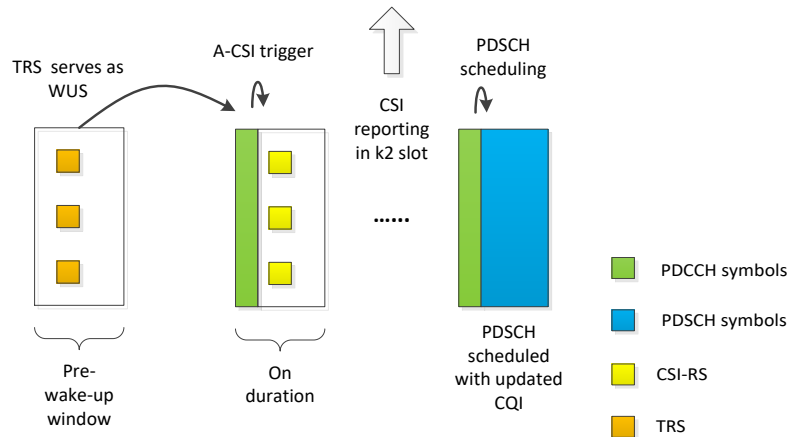


Figure 11: Single TRS Signaling in PWU Window

Observation 25: The use of single or separate transmissions of CSI-RS and WUS in the PWU window offers varying tradeoffs in terms of resource efficiency, scheduling flexibility and robustness.

3.3.2 A-CSI triggering and DRX inactivity timer

A-CSI trigger should be issued at the beginning of ON duration, so that gNB can get CQI report as early as possible in the case there is data to be scheduled to the UE. Even for the case there is no data to be scheduled for this cycle, it may improve efficiency to perform A-CSI measurement and reporting so that gNB has updated CQI for subsequent cycle. However, in Rel-15, there is currently a restriction that CSI measurement cannot happen outside the active time. Also, generally, A-CSI trigger does not start DRX inactivity timer. This means to support such A-CSI operation, ON duration cannot be configured to be too short, but this would be wasteful of power because for the DRX cycles without A-CSI triggering, UE would be suffering from such long ON duration regardless. To address such issue, we have the following proposal:

Proposal 19: When UE receives A-CSI trigger during ON duration, it always performs CSI measurement and reporting even if it requires going beyond the active time.

3.3.3 Dynamic adaptation of DRX configuration

In [10], it was agreed to study dynamic adaptation of DRX configuration as a candidate DRX enhancement technique. The principle of this technique is dynamically changing the DRX configuration parameters according to the short-term time-varying traffic condition. However, there are observations that may impair the usefulness of this technique.

First, the main benefit of DRX lies in putting the entire modem system in the deep sleep state. For relatively short sleep durations, the sleep power may not be as low as that of the deep sleep state. Also, frequent transition between sleep and active states may affect adversely due to transition overhead.

Second, all the timer-based processes of DRX are optimized for relatively long inter-arrival time. However, with a short inter-arrival time, the packets tend to first get accumulated in the buffer before being served to the UE through the physical layer. Therefore, the scheduler has a control on the physical-layer serving pattern of the traffic, and it may not be completely random. In this case, direct indication of sleep from the gNB may be more efficient than the timer-based DRX operation.

Observation 26: DRX operation based on timers may not be adequate for traffics with a short inter-arrival time.

Based on the above observation, for short-term power adaptation, we believe that direct signaling from gNB (e.g., by DCI) initiating UE power state change may be better option than DRX configuration parameter adaptation. DRX can still be applied for a larger time scale, where conventional semi-static RRC-configuration may be enough for parameter adaptation.

The topic of DRX configuration parameter adaptation is also discussed in our companion contribution submitted to RAN2 [18]. We think that RAN2 should be the primary venue for discussing this topic.

3.4 Adaptation to achieve reducing PDCCH monitoring/decoding

3.4.1 Semi-persistent search space configuration

In previous studies, for example, in [1]-[7], it was observed that PDCCH blind decoding without grant consumes a significant portion of energy in RRC_CONNECTED mode, and thus it is desirable to adapt the resources for PDCCH monitoring according to the traffic load, to minimize the chance of UE trying blind decoding in vain. In the current Rel-15 framework, adapting PDCCH resources may not be straightforward and include some issues. For example, we can rely on RRC re-configuration of search space sets, which would take tens of milliseconds with ambiguous timing. Another option is utilizing multiple BWPs with different search space configurations: for example, BWP1 may be configured with small SS periodicities, while BWP2 is configured with large SS periodicities. Switching between these two BWPs can effectively adapt the PDCCH resources to the traffic condition in a short latency. However, BWP is rather a scarce resource (only 4 BWPs per CC) with many other important applications and thus the adaptation flexibility will be restricted.

Based on the above arguments, the most suitable option for search space adaptation may be DCI-based signaling. For this purpose, we may define two states of a search space set, active and inactive states. When the SS set is in the active state, it is treated the same as a Rel-15 SS set, and UE monitors PDCCH within the set. However, when the SS set is in the inactive state, then the UE refrains from monitoring the set.

Once a SS set is configured, it may be inactive by default. Later, if the UE receives an activation DCI command for the SS set, it switches to the active state and the UE starts monitoring the SS set. The DCI command may be delivered from an *anchor SS set*, which is configured to be active at all times. The activation of the SS set may be semi-persistent, and the SS set may switch back to the inactive state based on another deactivation DCI command or on the expiry of an inactivity timer.

Note that the above described scheme can be applied both to single carrier and CA scenarios. In the single carrier case, at least one SS set in a BWP may be the anchor SS set and deliver activation/deactivation commands for the other SS sets in the same BWP. In the CA scenario, the anchor SS set can be configured on PCell or PSCell and deliver activation/deactivation commands for the other SS sets on the same or different CCs. A simple example of DCI-based SS set activation/deactivation in the CA scenario is shown in Figure 12.

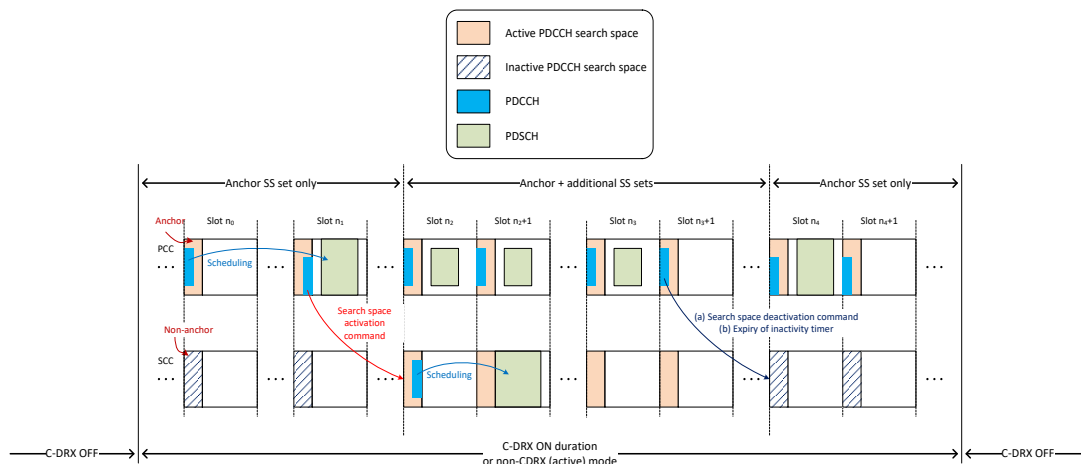


Figure 12: Search space set activation/deactivation operation

Proposal 20: DCI-based signaling is used for dynamic activation/deactivation of configured SS sets in PCell, PScell, and SCell(s).

3.4.2 PDCCH occasion skipping

In previous studies [7], [8], the statistics of the time gap (denoted by “scheduling gap” hereafter) between two consecutive scheduling opportunities for a UE in active time was investigated. The common observation from the previous studies is that the scheduling pattern for a UE is clustered in the time-domain and, from the UE’s perspective, there is a long vacancy of scheduling between two adjacent clusters.

During the scheduling gap, there is no PDSCH/PUSCH scheduled, and thus it is likely that the PDCCH monitoring occasions do not contain any scheduling grant. Therefore, if the gap can be indicated to the UE in advance, the UE can save power by skipping some PDCCH monitoring and switch to the sleep state. A relevant assumption is that the statistics of the gap is in a relatively small-time scale and may not be efficiently handled by C-DRX operation, which usually takes care of inactivity of relatively large time scale, as we discussed in Section 3.3.3. Therefore, most of the power saving during the gap may come from enhancing the opportunity of microsleep.

Observation 27: Considering the dynamic behavior of the scheduling gap, a possible option for the gap indication may be DCI signaling.

“Go-to-sleep” signaling that puts a UE back to DRX state has been proposed as a potential technique for UE power saving [10]. In its most basic form, it can be similar to the Rel-15 MAC-CE based DRX command, but transmitted over L1 signaling. It should be evident that such a “go-to-sleep” signal can be enhanced to indicate the length of the scheduling gap, and going to DRX state is just one of the options that can be indicated.

Observation 28: “Go-to-sleep” signaling, if introduced, should be capable of indicating the scheduling gap, in addition to the option of indicating transition to DRX state. Otherwise, there would not be enough enhancement over existing DRX command already supported in Rel-15.

Furthermore, in some respects, the indication is related to a specific format of slots without any scheduling grant and PDSCH/PUSCH. Therefore, a simple approach will be extending the existing Rel-15 SFI framework. For example, on top of the existing symbol states in Rel-15, such as downlink, uplink, and flexible, we can add a new state (which is called a “Null” state in Table 1) to indicate the UE that it is not required to transmit and receive any channels (PDCCH and PDSCH/PUSCH) on the corresponding symbol. Also, although a further study is needed, the null state may precede any other states (D, U, or F) and override semi-statically configured slot formats. Some example overriding rules between states are shown in Table 2.

Table 1: Symbol states

State	DL	UL	Flexible	Null
Application	Downlink	Uplink	DL or UL	No signal

Table 2: Overriding rules (O: dynamic overrides semi-static, X: overriding not allowed)

State		Semi-static resources						
		D	U	F	SSB	PRACH	Type0 CSS	Null
Dynamic (SFI)	D	O	X	O	O	X	O	X
	U	X	O	O	X	O	X	X
	F	X	X	O	O	O	O	X
	N	O	O	O	O	O	O	O

For the details of signaling, we can rely on the existing DCI format 2₀. Note that, in Rel-15, only 56 out of 256 entries in the slot format table (Table 11.1.1-1 in TS 38.213) are occupied. The remaining rows of the table can thus be populated with new slot formats with different null-symbol combinations. One caveat is that Rel-15 SFI is based on group-common signaling, while the indication of the scheduling gap is UE-specific in nature. Even in that case, the same DCI format and configuration may be re-used for signaling, for example, in a UE-specific SS set or with a UE-specific RNTI.

Proposal 21: Slot format indication in Rel-15 can be enhanced for the indication of “no signal” occasions, which the UE can utilize for power saving operations, e.g., PDCCH occasion skipping.

To support this proposal, some power consumption evaluation results are provided in [15].

3.5 Adaptation to number of antennas

3.5.1 BWP-based antenna adaptation

Adapting the number of antennas provides a tradeoff between performance and power consumption. That said, there are many scenarios for which operation with fewer number of antennas does not adversely affect performance (or effect is negligible), while the power saving benefit remains intact. For example:

- When the number of Rx antennas is greater than the rank supported by the channel, and the channel conditions are good such that combining gain or beamforming gain from the additional Rx antenna does not have much benefit.
- When the SNR requirement for the physical layer channel (e.g. PDCCH) is lenient, and only single layer transmission is supported. The requirement can be easily met with reduced number of antennas (e.g. 2Rx for PDCCH reception instead of 4Rx)

Figure 13 illustrates the performance of PDCCH decoding with 2 Tx antennas, 2 and 4 Rx antennas over the TDL-A (30ns delay spread) and TDL-C (300ns delay spread) channels. It can be observed that while reducing the number of Rx antennas affects performance, the change can be largely negated by increasing the aggregation level from 8 to 16 with the reduced number of antennas.

Observation 29: gNB can increase PDCCH aggregation level to compensate for PDCCH performance loss which may occur if the UE reduces the number of Rx antennas to save power.

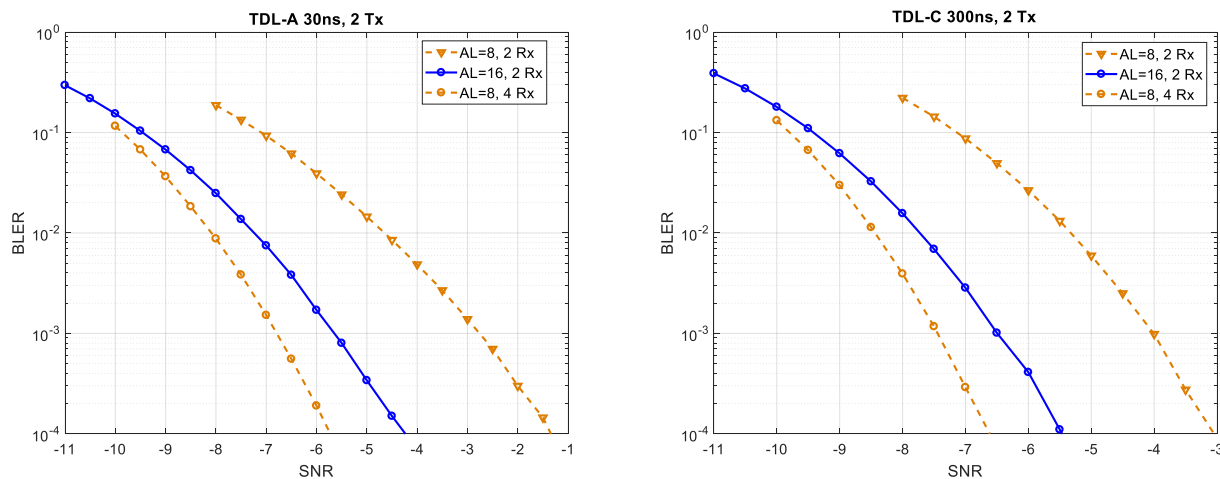


Figure 13: Performance of PDCCH with 2 and 4 Rx antennas.

Regarding the data channels, in Rel-15, the number of MIMO layers is RRC configured per cell for downlink transmissions (*maxMIMO-Layers* in *PDSCH-ServingCellConfig*) and both per BWP (*maxLayers* in *pusch-Config* for codebook-based PUSCH) and per cell (*maxMIMO-Layers* in *PUSCH-ServingCellConfig* for codebook and non-codebook based PUSCH) for uplink transmissions [12]. RAN-1 concluded that a per-BWP parameter indicating maximum rank for PDSCH is needed [13],[14].

The per-cell maximum rank parameter enables a form of antenna adaptation for PDSCH transmissions. However, since it is RRC-based, it suffers from the high latency of RRC reconfiguration and associated ambiguity duration.

Rel-15 BWP adaptation offers a framework that potentially fits very well with antenna adaptation. To save power with BWP adaptation, the UE may by default operate in a “low power BWP” and only switch to a “high performance” BWP during periods of high traffic activities. It would make sense to operate with fewer number of Rx antennas for the low power BWP, as it is intended mostly for PDCCH monitoring and reception of small data over PDSCH.

Based on above discussion, adding support for per-BWP maximum for PDSCH in Rel-16, as is done for codebook based PUSCH in Rel-15, provides the best combination of low latency and flexibility. It should be noted that, in Rel-15, many of the MIMO parameters already have per-BWP configuration.

Proposal 22: The RAN1 agreement of introducing a per-BWP RRC parameter indicating maximum rank for PDSCH should be adopted in Rel-16.

In this framework, switching BWP also potentially adapts the number of antennas. During BWP transition, RF transceiver may have to undergo some reconfiguration and it should be able to accommodate enabling/disabling of other antenna/RF chains. If the BWP transition time already defined is not sufficient, new transition time with larger delay can be specified.

Substantial power reduction is observed between PDCCH monitoring with 2-Rx antenna over narrow BW with $k_0 > 0$, and the baseline of PDCCH monitoring with 4-Rx antenna over full BW with $k_0 = 0$. According to the power model agreed in #94bis, reducing BW from 100MHz to 20MHz can reduce power by 60%, and independently, reducing the number of antennas from 4Rx to 2Rx can reduce power by 30%. It is logical to assume that the combination of reduced BW and reduced number of antennas can achieve power saving somewhere between 60% to 70%.

3.5.2 Multi-panel Designs for FR2

One aspect where antenna adaptation can be very effective for power saving is multiple panel operations. Especially for mmW systems with multiple antenna array modules (panels), each antenna module may largely be composed of analog components, such as filters, PLLs, and power amplifiers, and the power consumption for each active antenna module can be very high. Therefore, UE may not want to keep all antenna modules on all the time. Instead, to save power, UE may only keep one or several antenna modules active, while keeping the others in sleep. However, UE is required to switch beams around to maintain links to gNBs, and beam switching can either be intra-module or inter-module. In general, there is a longer time delay for inter-module beam switching than the intra-module case, especially when the target module first needs to be awakened from the sleep state. This large delay can be an issue when beam switching is required in aperiodic and unplanned manner: When DCI triggers UE's beam switching (for example, reception of aperiodic CSI-RS for P3 beam management, TCI indication for PDSCH, and SRS beam indication in DCI), the time offset between the DCI and the first symbol of the new beam should be large enough to accommodate the inter-module beam switching delay.

In Rel-15, beam switching timing in FR2 has been discussed and some candidate delay values have been agreed in the UE feature list (e.g., FG 2-2 and FG 2-28). In the upcoming discussion in Rel-16, it should be clarified whether both intra- and inter-module beam switching with power saving has been accounted for in the candidate delay values.

Proposal 23: Practical issues with multi-panel power-saving in FR2, such as inter-panel beam switching latency, should be identified in the power-saving study.

3.6 UE Assistance Information

In [10], it is recommended that the following power saving schemes based on UE assistance information are to be further studied:

- *UE assistance information/feedback to assist network in configurations for UE adaptation*
 - *UE preferred processing timeline parameters, e.g., K_0 , K_1 , K_2 values*
 - *UE preferred BWP information/configuration*
 - *UE preferred antenna configuration, including MIMO layers, antenna panel awareness information*
 - *UE assistance/feedback on the DRX configurations/parameters*
 - *UE preferred BWP provided to assist network in BWP switching*
 - *UE request on SCell/SCG activation/de-activation/configuration*
 - *UE preferred PDCCH monitoring parameters/search space configuration/maximum number of blind decoding*

In LTE, some UEAssistanceInformation IEs can be used by the UE to inform the network of its power status. For example, powerPrefIndication IE can be used to indicate whether the UE prefers power saving mode, although the indication alone does not contain enough information for re-configuring the UE for power-saving operations; it is up to gNBs implementation. In Rel-15, NR has already adopted some UEAssistanceInformation IEs from LTE, such as

delayBudgetReport and overheatingAssistance. Undoubtedly, these are far from enough to provide the UE with enhanced power saving gain, and there are many other configuration parameters and features to be considered, as captured in [10].

Proposal 24: In addition to the UE assistance information listed in [10], the following information can be added:

- Number of UL/DL carriers
- TDD slot format
- UE mobility
- Preferred CC grouping

Like LTE, the primary goal of power saving UE assistance information is to provide a mechanism for a UE to endorse the network to change the configuration related to the UE's power consumption. This is a direct request for a specific parameter change, while the acceptance is up to gNB. In Rel-16, in connection to all the discussion on dynamic power saving schemes in the previous sections, the second role of UE assistance information may be conveying UE's sets of preferred parameters to the network. In other words, in most cases, the dynamic power saving schemes can either be "selection among multiple entities" (e.g., BWP switching) or "activation/deactivation of an entity" (e.g., SCell activation/deactivation), while the UE assistance information can be the foundation for configuring the one or more entities (e.g., power saving BWP, PDCCH monitoring configuration, etc.). As the preferred configuration and extent of power saving may be very different across UEs, it would be reasonable that the UE informs the gNB with its preferred sets of configuration parameters, at least in a semi-static manner.

Proposal 25: Rel-16 UE assistance information can be used to convey UE's preferred parameter sets, to facilitate the configuration related to dynamic power saving operations.

4 Conclusion

Observation 1: Two-stage wake-up is the key idea for PDCCH-based WUS to facilitate power saving for DRX.

Observation 2: PDCCH can be studied as a candidate WUS waveform in Rel-16. Two strategies of introducing PDCCH-WUS in Rel-16 can be considered:

Observation 3: Rel-15 PDCCH configuration and multiplexing rules can be leveraged for PDCCH-WUS.

Observation 4: PDCCH-WUS resource utilization per UE or UE group is determined by the aggregation level set by the gNB.

Observation 5: PDCCH-WUS meets the misdetection rate criterion and provides extremely low false alarm rate.

Observation 6: DCI in PDCCH-WUS can carry information in addition to a binary wake-up indication. For example, a target BWP ID, A-CSI trigger, and a fine-granularity UE wake-up indication.

Observation 7: The UE can use the DMRS sequence to detect PDCCH-WUS presence prior to decoding DCI, reducing complexity and avoiding unnecessary decoding attempts.

Observation 8: CSI-RS-type waveform can perform as both WUS and warming-up resources.

Observation 9: Rel-15 CSI-RS configuration and multiplexing rules can be leveraged for CSI-RS-WUS.

Observation 10: For FR2, CSI-RS-WUS with beam sweeping may be more well-suited than PDCCH-WUS.

Observation 11: If PDCCH-based or CSI-RS-based WUS can deliver significant awake energy reduction for detection compared to normalized baseline of ON duration, the gain for further overall power saving may diminish even if another WUS scheme can achieve much lower awake energy for detection.

Observation 12: Discussion on signaling that triggers a power adaptation technique should be coupled with the discussion on the power adaptation technique. A unified signaling scheme may be undesirable.

Observation 13: Not monitoring PDCCH on SCell is a necessary but not a sufficient condition for improving SCell energy efficiency. Other periodic activities including CSI measurement and reporting, beam management, and/or RRM may still be performed at much reduced duty cycle. There is tradeoff between minimizing SCell power consumption level and maintaining sufficient state information about the SCell such that UE can very quickly transition the SCell to activation state when traffic arrives.

Observation 14: During Rel-15 standardization of BWP adaptation, there has been extensive and thorough discussion on whether CSI measurement in non-active BWP should be supported. In the end, the only such supported case is for RRM gap-based measurement. The decision was motivated by BWP adaptation for power saving use case and there is no need to reopen the discussion.

Observation 15: Rel-15 does not fully support triggering A-CSI request and BWP switch in the same slot. In some cases, A-CSI is delayed, and traffic performance may degrade, resulting in effectively higher BWP transition overhead.

Observation 16: It is beneficial to have a L1 signal that can trigger BWP switch and at the same time trigger A-CSI measurement and reporting. Given that DCI signaling is already supported for BWP switching in Rel-15, that can be used as the starting point.

Observation 17: A group of cells sharing the same wideband RF transceiver may be able to exercise more microsleep if gNB scheduler can align unscheduled slots across cells.

Observation 18: For the power saving BWP used for DRX wake-up, large minimum DL and UL scheduling offset facilitates additional power saving. Minimum DL and UL scheduling offset should be configured to be similar in value.

Observation 19: Ensuring a gap between PDCCH reception and PDSCH transmission in TDRA table configuration for same slot scheduling may have some drawbacks compared to the baseline of full-slot, back-to-back transmissions

Observation 20: Compared to same-slot scheduling with TDRA table restriction, cross-slot scheduling for UE power saving has wider applicability and does not rely on any additional signalling from the network.

Observation 21: The amount of UE power saving achievable based on same-slot scheduling with TDRA table restriction is highly dependent on UE implementation, whereas for cross-slot scheduling, the dependency is more relaxed leading to more robust gain.

Observation 22: TDRA entry selection does not resolve the A-CSI triggering offset issue.

Observation 23: To guarantee $k_0 \geq x$ using TDRA entry selection with multiple configured BWP, update to TDRA entry selection for the non-active BWPs may be needed, leading to extra signalling overhead.

Observation 24: CSI-RS and WUS Transmissions in the PWU has the potential to provide significant UE power savings.

Observation 25: The use of single or separate transmissions of CSI-RS and WUS in the PWU window offers varying tradeoffs in terms of resource efficiency, scheduling flexibility and robustness.

Observation 26: DRX operation based on timers may not be adequate for traffics with a short inter-arrival time.

Observation 27: Considering the dynamic behavior of the scheduling gap, a possible option for the gap indication may be DCI signaling.

Observation 28: "Go-to-sleep" signaling, if introduced, should be capable of indicating the scheduling gap, in addition to the option of indicating transition to DRX state. Otherwise, there would not be enough enhancement over existing DRX command already supported in Rel-15.

Observation 29: gNB can increase PDCCH aggregation level to compensate for PDCCH performance loss which may occur if the UE reduces the number of Rx antennas to save power.

Proposal 1: For the study of PDCCH as a candidate WUS waveform, designing a new dedicated PDCCH-WUS leveraging Rel-15 PDCCH should be the preferred approach.

Proposal 2: Flexible time-domain location of a search space can be considered as a Rel-16 enhancement for PDCCH-WUS configuration.

Proposal 3: CSI-RS can be studied as a candidate WUS waveform in Rel-16.

Proposal 4: At least in FR2, beam sweeping can be applied for WUS transmission.

Proposal 5: SCell dormancy state should be supported in Rel-16 NR for power saving. The BWP framework should be enhanced to support operation with "dormant BWP".

Proposal 6: Semi-persistent search space configuration based on DCI-signaling can be considered to support fast transitioning in and out of SCell dormancy state.

Proposal 7: A-TRS/CSI triggering from SCell activation command should be supported in Rel-16 to greatly reduce SCell activation time.

Proposal 8: DCI-based signaling for SCell activation/deactivation should be supported in Rel-16, if other enhancements are also adopted to greatly reduce the overall activation timeline.

Proposal 9: Cross-BWP triggering of A-CSI measurement and reporting should be supported in Rel-16. Triggering of aperiodic CSI and BWP switch in the same slot should be fully functional.

Proposal 10: One of the configured BWPs can be designated as the BWP which is always activated when UE starts for DRX ON duration. Other association between BWP and DRX operation can be considered for Rel-16.

Proposal 11: Serving cells with different uses can be assigned to different groups, and different cell groups can be configured with different DRX parameters. gNB can time-align scheduling across cells within a group to maximize microsleep. UE may recommend grouping of cells via UE assistance information feedback.

Proposal 12: For CSI acquisition, allow non-zero aperiodic CSI-RS triggering offset in the case that all the associated trigger states do not contain QCL Type D information.

Proposal 13: NR should consider explicit configuration/signaling of minimum DL scheduling offset and minimum UL scheduling offset for Rel-16.

Proposal 14: Timing determination for A-CSI-RS relative to the triggering grant can be a function of the configured A-CSI triggering offset and the minimum DL scheduling offset.

Proposal 15: A-CSI triggering offset can be indicated in the DCI with the A-CSI request, and satisfy the minimum DL scheduling offset constraint.

Proposal 16: Minimum DL scheduling offset or minimum UL scheduling offset configuration is per DL or UL BWP respectively, and RRC-configured as part of the BWP configuration.

Proposal 17: Minimum scheduling offset determination should be based on UE capability signaling and/or UE-assistance framework.

Proposal 18: Pre-wake-up window is considered as a C-DRX enhancement scheme for power saving and performance improvement.

Proposal 19: When UE receives A-CSI trigger during ON duration, it always performs CSI measurement and reporting even if it requires going beyond the active time.

Proposal 20: DCI-based signaling is used for dynamic activation/deactivation of configured SS sets in PCell, PScell, and SCell(s).

Proposal 21: Slot format indication in Rel-15 can be enhanced for the indication of “no signal” occasions, which the UE can utilize for power saving operations, e.g., PDCCH occasion skipping.

Proposal 22: The RAN1 agreement of introducing a per-BWP RRC parameter indicating maximum rank for PDSCH should be adopted in Rel-16.

Proposal 23: Practical issues with multi-panel power-saving in FR2, such as inter-panel beam switching latency, should be identified in the power-saving study.

Proposal 24: In addition to the UE assistance information listed in [10], the following information can be added: -
Number of UL/DL carriers - TDD slot format - UE mobility - Preferred CC grouping

Proposal 25: Rel-16 UE assistance information can be used to convey UE’s preferred parameter sets, to facilitate the configuration related to dynamic power saving operations.

5 References

- [1] R1-1813448, “Triggering Adaptation of UE Power Consumption Characteristics”, RAN1#95, Qualcomm.
- [2] R1-1900912, “Triggering Adaptation of UE Power Consumption Characteristics”, RAN1-AH-1901, Qualcomm.
- [3] R1-1716444, “Performance evaluation for distributed CRC,” RAN1-AH-1703, Qualcomm Inc.
- [4] R1-166368, “UE Power Consideration based on Days-of-Use”, RAN1#86, Qualcomm.
- [5] R1-1609557, “On Techniques for Enhanced UE Power Efficiency”, RAN1#86bis, MediaTek.
- [6] RP-172318, “Motivation for Study on UE Power Saving in NR”, RAN#78, VIVO.
- [7] R1-1809333, “General Consideration on UE Power Saving in Rel-16”, RAN1#94, Huawei, HiSilicon.
- [8] R1-1813447, “UE Adaptation to the Traffic and UE Power Consumption Characteristics”, RAN1#95, Qualcomm.
- [9] R2-1714289, “Running CR for euCA Stage-2”
- [10] R1-1814387, “Study on UE Power Saving (Release 16)”, draft TR 38.840.
- [11] R4-1811913, “CR to TS 38.133: Implementation of endorsed draft CRs from RAN4-AH-1807 and RAN4 #88”, Intel.
- [12] TS 38.331 v.15.4.0, Radio Resource Control (RRC) protocol specification (Release 15).
- [13] R1-1813952, “Reply LS on MIMO layer configuration,” RAN1.
- [14] RAN1 #95 Chairman’s notes.
- [15] R1-1903015, “Evaluation Methodology and Results for UE Power Saving”, RAN1 #96, Qualcomm.
- [16] R1- 1903028, “Fast SCG and Scell activation”, RAN1 #96, Qualcomm
- [17] R1- 1903026, “Cross-Carrier Scheduling with Different Numerologies”, RAN1 #96, Qualcomm
- [18] R2-1901764, “DRX enhancements for better adaptation to traffic”, RAN2 #105, Qualcomm
- [19] R2-1901766, “DRX group for carrier aggregation”, RAN2 #105, Qualcomm