GTI 5G Device Power Consumption White Paper

GTI 5G Device Power Consumption White Paper

Version:	V2.0
Deliverable Type	□ **Procedural Document** √ **Working Document**
Confidential Level	√ **Open to GTI Operator Members** √ **Open to GTI Partners** □ **Open to Public**
Program	5G eMBB
Working Group	Terminal WG
Project	Project 3: New Device
Task	5G device Power Consumption Task Force
Source members	China Mobile, VIVO, MTK, Smarter micro, Qorvo, OPPO, Anritsu, Keysight, XIAOMI, Intel
Support members	Quectel, Unisoc, SIMCom,Starpoint,Datang Mobile, Taiyo Yuden, Qualcomm, Rohde-Schwarz.
Editor	ZhangYuan, Simon George-Kelso, Francesc Boixadera, Hu Hao, Jinming Zhao, Guiying Wang, YangyangPeng, Lawrence.Tao, Xiaowei Jiang, Luting Kong, Jiang Bo, Zhihua Shi, Jianwei Zhang, TT Chiang, Shaoliang Tang.
Last Edit Date	03-06-2019
Approval Date	03-06-2019

Document History

Date	Meeting #	Version #	Revision Contents
21-01-2019	24th GTI Workshop	V1.0	The first version of GTI 5G Device Power Consumption Whitepaper.
03-06-2019	25th GTI Workshop	V2.0	Update 5.3 parameter configuration recommendation
19-06-2019	25th GTI Workshop	V2.01	Formatting edits

Table of Contents

1 Executive Summary

5G brings wider bandwidth, higher data rate and lower delay, but it also brings new challenges to the power consumption of 5G devices. This whitepaper provides the analysis of the factors of power consumption, the test solution and the performance requirements of power consumption for 5G device.

First of all, the power consumption of the components that have the greatest impact on terminal power consumption is introduced, including RFFE (such as PA, AD/DA, filter), RFIC, BP, AP and screen parts, and analyze the power consumption performance of the main solutions in the current 5G component industry. Secondly, based on the 5G technical features, the possible power consumption problems (such as wider bandwidth, uplink dual transmission, etc.) are discussed, and the 5G power saving schemes (such as BWP, cross-slot scheduling, DRX, periodical PDCCH monitoring). At the same time, based on the analysis of the power consumption performance of the components, the influence of 5G technical features on power consumption is deduced preliminarily. Eliminate the unnecessary technical requirements which seriously affect the device power consumption and have no obvious benefit. Summarize the key technical characteristics that affect the 5G device power consumption, pending further test and verification. Thirdly, based on the analysis results of typical service types and user models, as well as the key technical characteristics affecting 5G terminal power consumption, we will delimit the test scope and complete the test cases. Discuss the specific test scheme with the instrument manufacturer. The instrument manufacturer is responsible for the test case development and recommends the power test scheme suitable for 5G terminal. Fourthly, to build the 5G terminal test environment. Based on the previous 5G terminal test scheme, the key technical features affecting the power consumption of 5G terminal are further tested and verified. Finally, the key technologies suitable for 5G terminal are selected. Finally, based on the test results, the power performance requirements of 5G terminal are finalized, including the power requirements of typical service types in RRC connection state and the power requirements of terminal in RRC idle state and RRC inactive state.

In the early stage, the analysis of power consumption for sub6G eMBB terminal and the vertical terminal is the main part, the start-up time of the research of power consumption for millimeter wave device depends on the industry maturity.

2 Abbreviations

Abbreviation	Explanation
3GPP	3rd Generation Partnership Project
AR/VR	augmented reality /virtual reality
AP	applications processor
BP	baseband processing
BWP	Bandwidth part
DL	Downlink
DRX	Discontinuous Reception
eMBB	Enhanced Mobile Broadband
eMBMS	Enhanced Multimedia broadcast multicast services
ETSI	European Telecommunication Standardisation Organisation
FCC	Federal Communications Commission
GTI	Global TD-LTE Initiative
HPUE	High power user equipment
IMT	International Mobile Telecommunication
ITU	International Telecommunication Union

ITU-R	International Telecommunication Union - Radio
LTE	Long Term Evolution
MNO	Mobile Network Operator
MWC	Mobile World Congress
MOOD	MBMS Operation On Demand
MIMO	Multiple input – Multiple output
NSA	Non-standalone
NR	New Radio
OAM	Operation, Administration and Maintenance
ODU	LTE outdoor CPE
OEM	Original Equipment Manufacturer
ODM	Original Device Manufacturer
QoS	Quality of Service
PDCCH	Physical downlink control channel
RAN	Radio Access Network
RRM	Radio Resource Management
RFFE	RF Front End
RFIC	Radio frequency integrated circuit
SA	Standalone
TD-LTE	Time Division Long Term Evolution
TDD	Time Division Duplex
TTI	Transmission time interval
UE	User equipment
UL	Uplink

3 Introduction

One of the 5G challenges is the power consumption of 5G devices. This whitepaper provides the analysis of the factors of power consumption, such as the key components /the 5G feature and the service type/the test solution and the performance requirements of power consumption for 5G device. In the early stage, the analysis of power consumption for sub6G eMBB terminal and the vertical terminal is the main part, the start-up time of the research of power consumption for millimeter wave device depends on the industry maturity.

4 Key components of power consumption

Comparing 5G and 4G, the 5G device needs to support new bands, wider bandwidth, more antennas, higher transmit power, new RAT and new services. The block diagram of the 5G device is shown in Figure 4-1. The new features might affect the power consumption of the blocks, such as antenna tuning unit, RFFE, RFIC, BP and AP. The factors of power consumption should be considered in the design and the development of the key components. The factors of power consumption of key components are analyzed and the trends of the solutions are discussed.

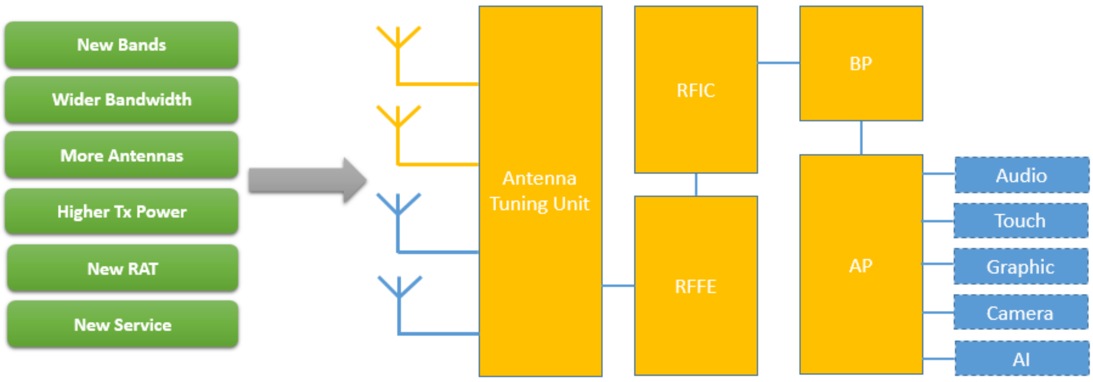


Figure 4-1 5G Device Block Diagram

The factors of power consumption for each component are listed in following in the aspect of 5G features. The change of the components power consumption is analyzed from the point of 5G feature.

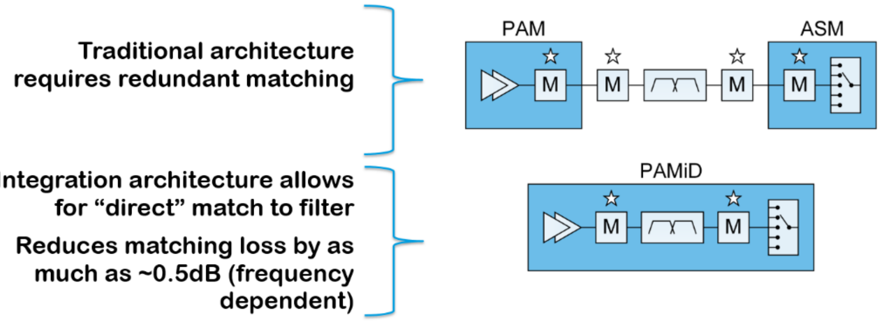
Table 4-1 Factors of Power Consumption - 5G Features vs. Key Component

5G Features	ATU	RFFE	RFIC	BP	AP
Bandwidth	√	√	√	√	√
UL-MIMO	√	√	√	√	√
HPUE	√	√	√	√	√
BWP	√	√	√	√	

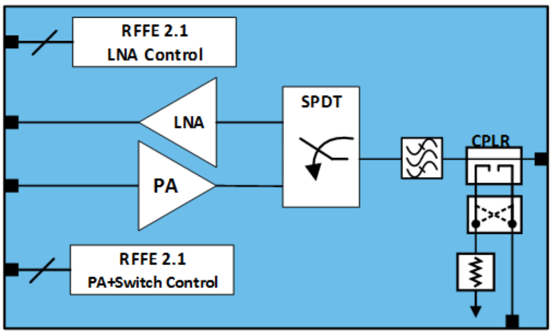
DRX				√	
Cross-slot scheduling				√	
Multiple PDCCH monitoring periodicities				√	
Measurement				√	
System information acquisition				√	
Paging				√	
RNA update				√	

4.1 RFFE

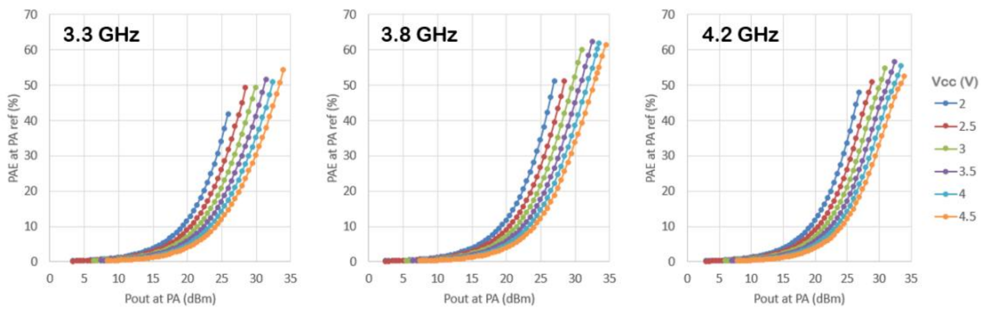
RF Front End Integration can reduce loss and improve power consumption



Integrated RF Front End as below is the best approach for 5G to save power consumption



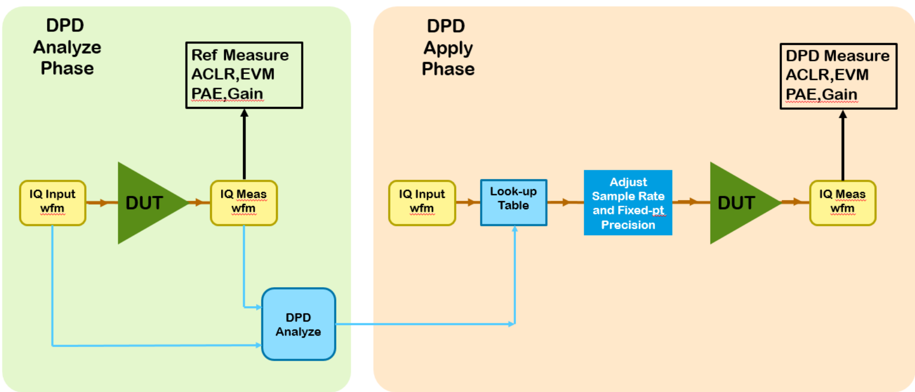
n77 PA efficacy inside the module.

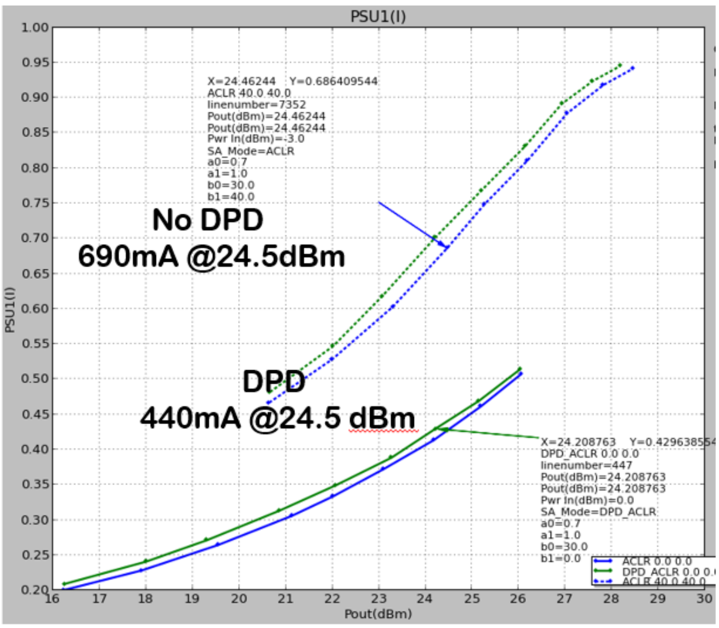


PA PAE [%] at Vcc=4V	3.3 GHz	3.8 GHz	4.2 GHz
Saturated PAE	51.5%	61.9%	55.4%
29.4 dBm (~6 dB back-off) 4G LTE Rel. 8	32.5 %	35.6 %	34.9 %
26.4 dBm (~9 dB back-off) 5G 64QAM, 100 MHz	19.2 %	20.1 %	19.8 %

Digital Predistortion Technology can further save power consumption.

DPD Operation





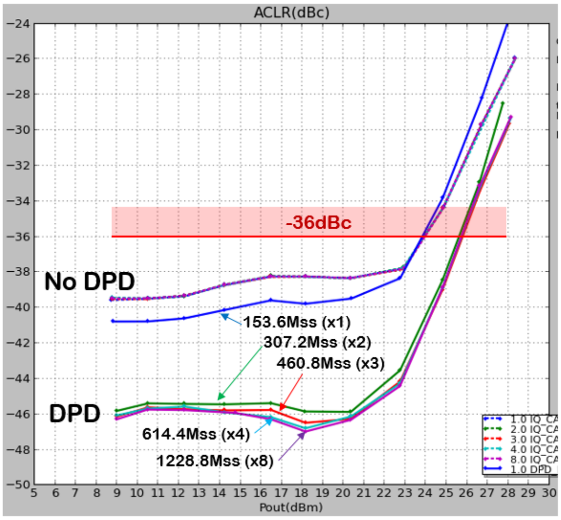
Simple memoryless DPD.

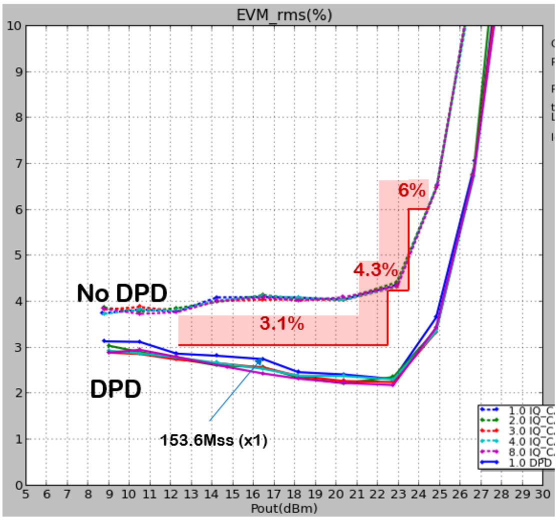
DPD enables dramatic power reduction

- 30% current reduction
- 30% PAE for 5G Waveform in APT mode.

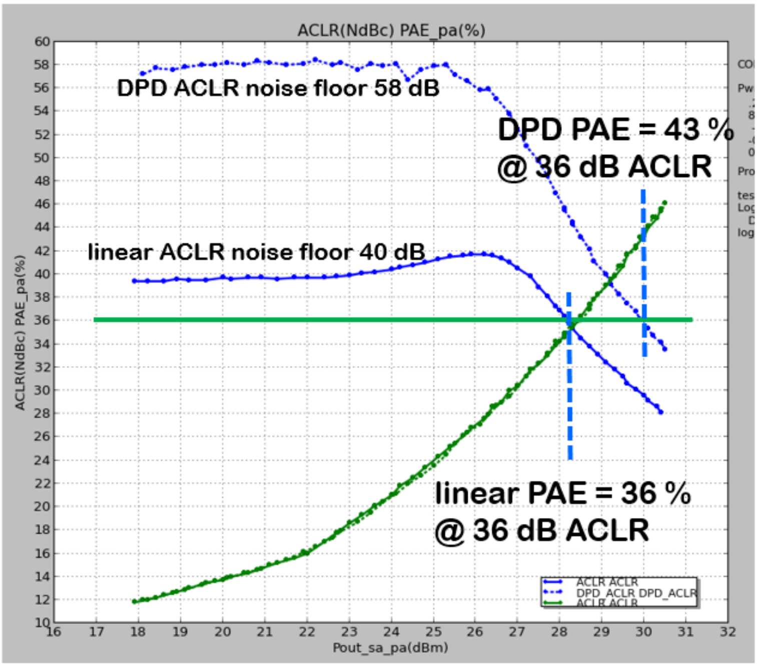
DPD increases linear power range by 1 dB.

Digital Predistortion Technology can also further improve linearity.





Digital Predistortion Technology improve both power consumption and linearity.



4.1.1 Basic Information of PA

The basic topology of power amplifier in handset devices can be shown as below.

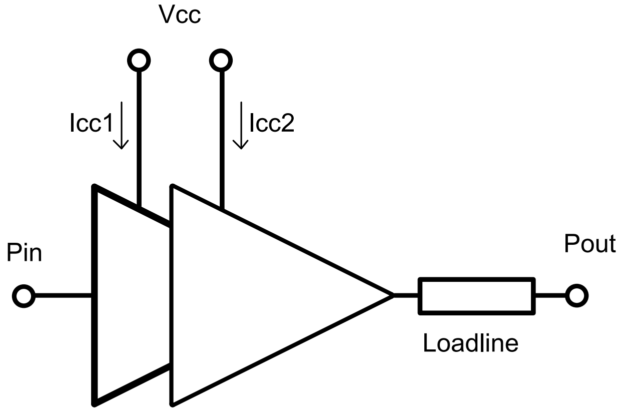


Figure 4-2 Basic topology of power amplifier

For the power amplifier, the efficiency (always noted as Power Add Efficiency or PAE) is always used to represent the power consumption character. The relationship between the efficiency and power consumption can be calculated as:

$$\text{PAE} = \frac{P_{\text{OUT}} - P_{IN}}{P_{dc}} = \frac{P_{\text{OUT}} - P_{IN}}{V_{dc} * I_{dc}}$$

There are many parameter related with the efficiency of the power amplifier, include back-off power, loadline loss, topology, etc. Before talking about how to improve the efficiency, we need to analyze the relationship between efficiency and other parameters.

For the loadline loss, the loss of the loadline consumes the power. The power consumed by the loadline in dBm equals to the number of the loadline loss in dB. The following table shows the relationship between the potential of the transmitted power in percentage and the loadline loss in dB.

Table 4-2 the relationship between transmitted power efficiency (%) and loadline loss (dB)

Load line loss (dB)	Transmitted Power Efficiency	Lost Power Efficiency
-0.1	98%	2%
-0.2	95%	5%
-0.3	93%	7%
-0.4	91%	9%
-0.5	89%	11%
-0.6	87%	13%

-0.7	85%	15%
-0.8	83%	17%
-0.9	81%	19%

For a typical loadline, the loss is between 0.3dB to 0.7dB, which means will degrade the PAE by 7 to 15%.
Another factor effects the efficiency of the power amplifier is the back-off power. For the topology of traditional linear power amplifiers like Class-A/Class-AB/Class-F, the relationship between the PAE and the output power can be shown in the following picture:

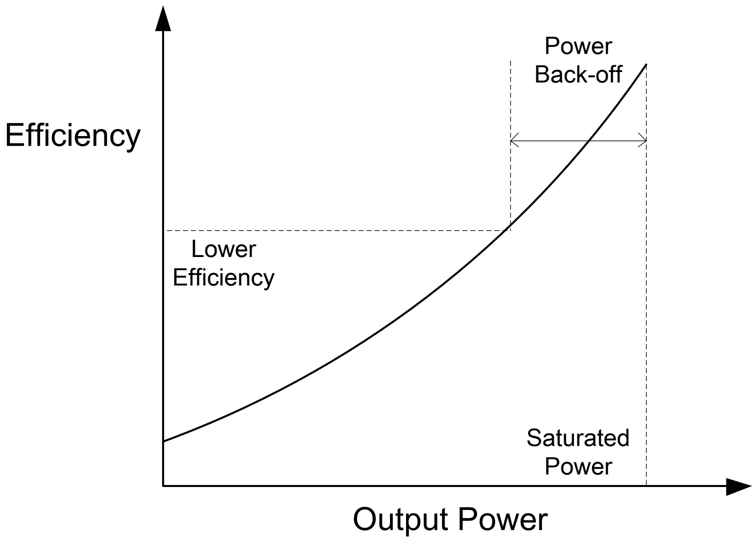


Figure 4-3 Typical relationship between efficiency and output power

As the picture shows, the PAE decreases with the power back-off from the peak power.

4.1.2 Methods of improving the PA efficiency

Based on the previous consideration, there are several methods of improving the efficiency.

- Average Power Tracking

Average Power Tracking (APT) is a technique that can be utilized for vary the supply voltage to a power amplifier on a timeslot basis in order to reduce power consumption of the PA.
With APT method, the supplied voltage of the power amplifier is adjusted according to output power level so that linearity of the power amplifier is maintained while the efficiency is improved.
But limited by the speed of the DC-DC converter, the APT voltage must be kept constant in the same time slot. Which means the efficiency cannot be improved if the power amplifier works at the highest power constantly.

- Envelope Tracking

The envelope tracking approach is one recommended power supply technique that maximizes the energy efficiency of the PA by keeping it in compression over the whole modulation cycle, instead of just at the peaks, by dynamically adjusting the supply voltage to the PA.
The ET technique was first developed in 1930s to handle the excessive energy consumption of

high-power amplitude modulation broadcast radio transmitters. Since constant amplitude ns and frequency modulation techniques displaced AM in the 1950s, ET became marginalized and irrelevant to engineering, where it languished as an academic curiosity. However, the increasing PAPR of modern signals with advanced digital modulation schemes such as 4G reinvigorate ET technique to achieve considerable energy saving in high-PAPR digital transmitters. More commercial ET Pas have been developed for modern 4G/5G wireless communications and beyond.

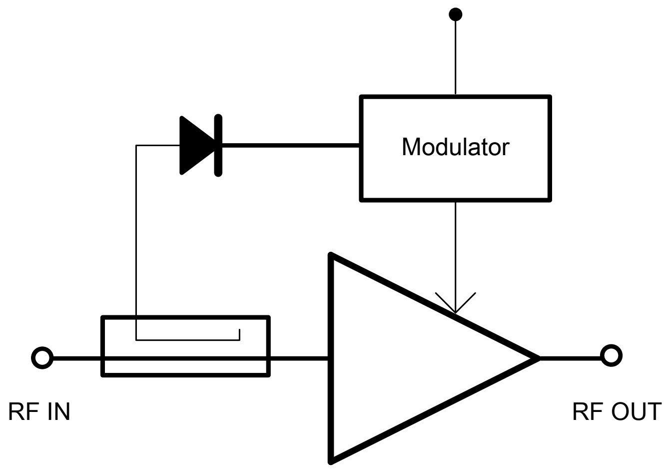


Figure 4-4 Envelope Tracking RF PA system

- High Voltage Supply

With higher supply voltage, the portion of knee voltage of the transistors can be minimized, which means the power amplifier will get higher efficiency.
The other advantage of high supply voltage is the loadline can also be increased. With the increasing of the loadline, the loadline loss will also be lower. This also can improve the efficiency. But because the supply voltage of the handset is 3.8V at nominal condition, another DC-DC boost is needed to get a high voltage. It should be noted that the DC-DC's efficiency is not 100%.

- Power Combination Techniques

The Power amplifiers usually need to support very large output power. So, power combination techniques are dispensable techniques to combine small power cells together.
Because the combiners are always at the last stage of the power amplifier, the efficiency of the combiners will affect the power amplifier's efficiency directly. There are many power combination techniques, includes voltage combination, current combination. No matter which combination techniques, low loss, high Q devices must be used.
With the combination techniques, the push-pull combiner has the advantage of doubling the impedance which as to be matched. This will lower the loss of the impedance transformation.

- The Doherty Amplifier

The Doherty amplifier was first proposed in 1963. The Doherty amplifier uses a configuration called active load-pull technique, which can modify the RF load of the power amplifier in different RF power levels. When the power back-offs from the peak power, the efficiency won't drop too much. Comparing with the traditional power amplifier topology, the power back-off efficiency can be improved.
In the Doherty amplifier includes two devices, which can be called "main" and "auxiliary" devices. The final maximum RF output power is the combined power of both devices. As the input drive level is reduced, both devices contribute to the output power until a certain point is reached, typically 6dB down from the maximum composite power, where the auxiliary amplifier shuts down and generates no more RF power; it is assumed that it will also cease to draw DC as well.

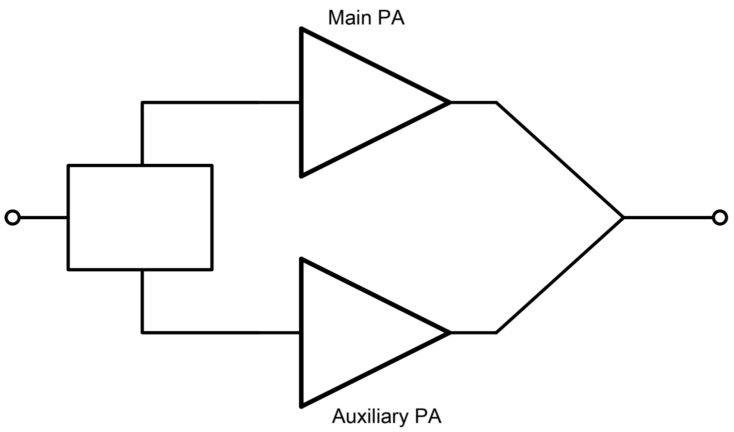


Figure 4-5 Basic topology of Doherty PA

In the Doherty amplifier, typical relationship between the power and efficiency can be shown as below. With the picture shows, the power back-off efficiency can be improved obviously.

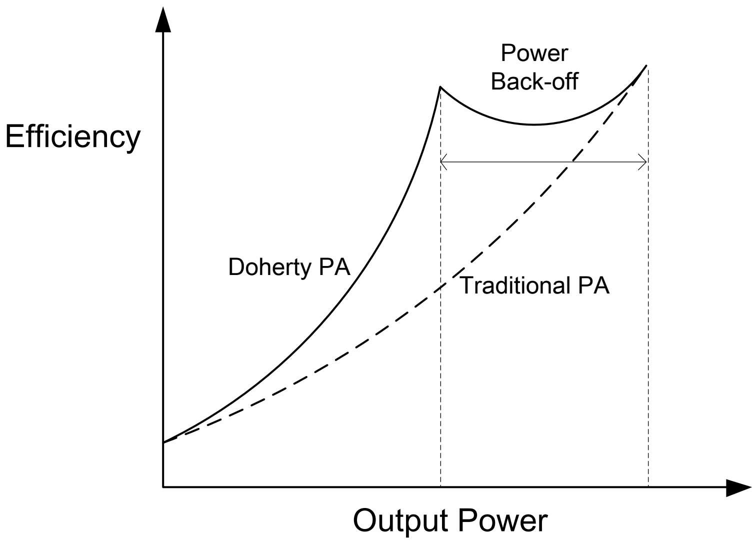


Figure 4-6 Efficiency of Doherty PA

The Doherty amplifier also has some disadvantages. First, the AM-AM and AM-PM of the amplifier cannot maintain constant as the auxiliary amplifier turning on and off with the changes of the input power. So, digital pre-distortion method is always adopted in the Doherty power amplifier. Second, the performance of the Doherty amplifier is very sensitive to the load. Based on those characters, Doherty amplifier topology is always adopted in the base station power amplifier design.

- The Reconfigurable Technique

The reconfigurable technique allows the power amplifier to adapt to different configuration in different bands and different modes. Smarter Micro has applied this technique into power amplifier design. With this technique, the loadline, bias point, and other configurations can be adjusted. Comparing with wide-band design, this technique will improve the efficiency at specified modes.

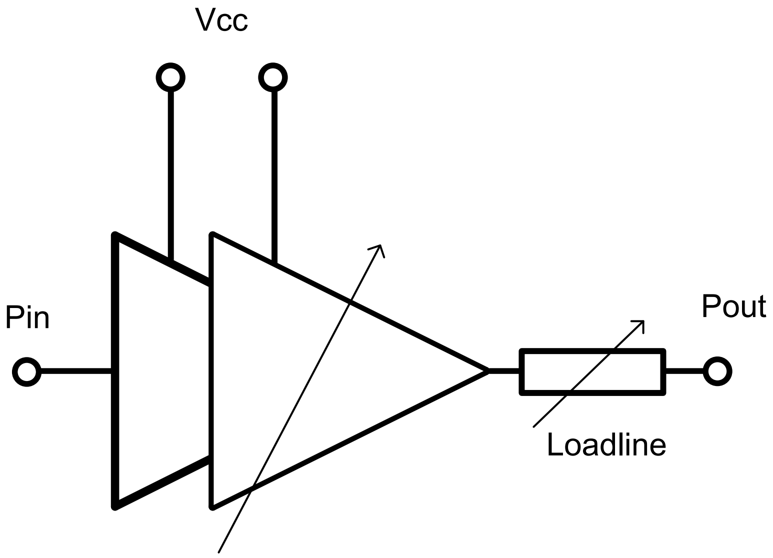


Figure 4-7 Topology of reconfigurable amplifier

4.2 BP

The baseband processing in an NR UE modem has to handle a significantly increased volume of data compared to its LTE counterpart, and this is reflected in increased power consumption at the highest data rates. In an earlier study [1], estimates were made of the likely baseband contribution to UE power consumption at peak throughput. These are reproduced below for reference.

Table 4-3 Estimated baseband processing power for different UE configurations

UE configuration	Power at peak throughput (FDD)	Power at peak throughput (TDD)
2x100MHz, 4x4 DL, 2x2 UL	4500mW	2970mW
1x100MHz, 4x4 DL, 2x2 UL	2250mW	1485mW

That study was based on a very simple set of basic assumptions – in practice a 2 carrier baseband does not require double the power of a single carrier baseband, as not all resources are duplicated. More recent results suggest that these estimates were reasonably accurate, being pessimistic for the 2 carrier UE, but optimistic for the 1 carrier UE.

However, power consumption at peak throughput is far from the whole story. Even in an active connection, a UE will spend a large proportion of its time monitoring the downlink control channel in TTIs which do not carry any data for it. It is therefore important that in this PDCCH-only state, power consumption is kept to a minimum.

This can be achieved by a combination of cross-slot scheduling, bandwidth part adaptation and MIMO restriction.

The UE can further reduce its average power consumption in connected mode (at the expense of higher latency) by entering a DRX cycle to reduce the time spent monitoring PDCCH. These techniques are discussed further in sections 5 and 6

4.3 AP

The applications processor in a smart phone comprises a number of processing cores and graphical processors which cooperate to support the computational requirements of the active applications. Multiple cores support a mixture of clock rates and processing capabilities, so power consumption can vary significantly depending on the number of cores that are active and the set of applications that are running. The highest processing loads are usually associated with display-intensive applications, and the combined power consumption of the applications processor and display can exceed 1000mW in some scenarios, even in flight mode.

AP power consumption can in many cases be considered independently from the UE modem power. Present day application data requirements are relatively modest in NR terms – even streaming of high definition video to the UE uses a comparatively small proportion of the data bandwidth that is available. However, the applications themselves can have a significant impact on modem power consumption. Interactive gaming applications can involve frequent transfer of small packets of data to give a real-time response, and if the update frequency falls within the inactivity timer period of the DRX cycle the modem will spend most of its time awake with increased power consumption. Updates from different applications, if their timing is not coordinated, increase the proportion of DRX cycles in which data is active and thereby increase power consumption.

With the increased bandwidth offered by NR it seems inevitable that new applications such as augmented reality gaming will combine intensive application processing with simultaneously high on-air transfer rates. Providing a good experience to mobile users who have to rely on finite battery capacity will require cooperation between application developers, UE manufacturers and network providers.

4.4 RFIC

The receivers and transmitters in the NR RFIC must operate over a wide range of frequencies, and a single implementation that will cover more than 2 decades (600MHz – 86GHz) of the RF spectrum is an unrealistic objective. Individual receivers and transmitters in the RFIC will be optimised for different sub-bands, and enabled as necessary for cell search and selection.

Power requirements will vary according to the frequency band and the number of active paths. In the earlier study [1], estimates of power consumption were made for a 100MHz FR1 carrier, and these are reproduced below in Table 4-4.

Table 4-4 Estimated transceiver power for a 100MHz FR1 carrier

	MIMO configuration	Active power
100MHz downlink	4x4	200-250mW
100MHz uplink	2x2	150-200mW

There is very little information so far in the public domain on FR2 power consumption, but higher losses at mmWave frequencies, higher bandwidth in the transceiver and more complex beamforming will all contribute to increased power consumption in this region of the spectrum.

Reducing the number of active receive or transmit paths will reduce power consumption, although the reduction is generally less than pro-rata – halving the number of paths might reduce power by around 30%.

Reducing the bandwidth and sampling rate can also reduce power in the transceiver, but the savings are generally more noticeable in the downlink, where they are accompanied by a

reduction in baseband processing power, than in the uplink where the PA power contribution is often dominant.
The most significant savings in transceiver power come from duty cycle reductions resulting from a good DRX configuration.

4.5 Screen

- Screen market status and trends

For electronics products, the material of the screen determines the display effect of the product to a large extent. If the screens are categorized by their material, the major screens of the smartphones can be divided by their material into two categories: one is LCD (Liquid Crystal Display) and the other is OLED (Organic Light-Emitting Diode). TFT and SLCD, which are more common in the market, belong to the category of LCD. Lightweight, flexible, higher resolution, larger size and other requirements will affect the future direction of mobile display industry.
The 4K resolution is defined as 4096 pixels in horizontal and the aspect ratio is still 16:9. There are two main types of 4K resolution: 4096 * 2160, which is widely used in the field of digital movies; 3840 * 2160, which is basically the 4K resolution we see in TV, displays, mobile phones and other daily devices. There are twice as many pixels for 3840 * 2160 in both horizontal and vertical directions comparing with the common FHD (1920 * 1080, 1080p). So the total pixel number is four times that of FHD. Since the number of pixels is four times that of FHD, driving a 4K screen can be a huge resource drain (especially GPU). When the screen's PPI (pixels per inch) is greater than 300, the human eye can't distinguish the pixels on the screen, such as retina screen (326 ppi). The 1080p screen (over 400 ppi) is clearer than the retina screen, so is 1440p (2560 * 1440, 2K). For 5.5 inch 4K, the number of ppi has reached an exaggerated 806. The display screen of VR application is much closer to the eye coupled with the lens effect and the left and right eye images are rendered separately. Even if the screen ppi reaches 400 or even 500, users can still feel the obvious sense of granularity. Thus, 4K device is a better choice. At a distance of 20 cm, the accuracy difference is negligible between 1080p and 2K. It is difficult to tell exactly what resolution it is by simply giving a small 1080p or 2K screen. Furthermore, high resolution means high power consumption, high processor load, which causes high power consumption and heat problem. Thus, many manufacturers are reluctant to use 2K screens. The higher screen resolution is only suitable for the mobile phone which supports VR function. For the general users, the accuracy of 1080p is enough.
Flexible screen refers to flexible OLED. It not only greatly benefits the manufacture of a new generation of smart phones, but also has a far-reaching impact on the application of wearable devices due to its low power consumption and flexibility. In the future, flexible screens will be widely used with the continuous penetration of personal smart devices. OLED uses plastic substrates, rather than common glass substrates, with the help of film packaging technology, and paste a protective film on the back of the panel, so that the panel becomes flexible, not easy to break. Flexible screens can be curled, but they cannot be folded. Currently on the market are ordinary rolled-up mobile phones since the shape of other parts of the mobile phone cannot be changed. In February 2016, Canadian researchers at Queens University's Human Media Laboratory unveiled the exciting invention of flexible screen technology, creating the world's first flexible-screen smartphone equipped with full-color, high-resolution displays and wireless technology. The phone, called ReFlex, comes with a 720p flexible OLED screen from GDisplay and Android 4.4 KitKat system. Because the battery and PCB are not yet bent, it only has a flexible screen with PCB and battery built below the screen. Many smart phone manufacturers show their interests in flexible screen for larger screens and smaller size.

- Factors of 5G device power consumption

The larger size and the higher resolution and brightness of the screen will cause more power consumption. The screen power consumption is calculated by its power, which is the product of current and voltage. When the dot pitch and the brightness of the screen remain unchanged, the larger the size and the higher the resolution. The screen power consumption increases with the number of luminous points. Similarly, when the brightness and the size of the screen main unchanged, the screen power consumption increases with the resolution. For 5G devices with larger screen sizes and 5G devices supporting VR, the problem of screen power consumption needs to be solved.

The power consumption depends on the content of video. When broadcasting advertisements or live video, white and color dominate, power consumption might be large. When playing simple graphics such as text, most of the screen is black, and the average power consumption is the lowest. As 5G terminal products support higher data rate and lower latency, and have broad prospects of video application, the proportion of video services of the typical user model will increase. We need to build a new typical service model for 5G user and further analyze the impact on 5G power consumption.

5 Factors of power consumption - 5G features

5.1 5.1 RRC idle state and RRC inactive state

5.1.1 Bandwidth

In NR, UE in RRC idle state or RRC inactive state can operate on an initial active DL BWP, which can be, e.g., 24RB, 48RB or 96RB configured by MIB for different SS/PBCH block and PDCCH subcarrier spacing [Section 13 in TS 38.213], to receive SSB, RMSI, paging, etc. The number of RBs for initial active DL BWP will impact on UE power consumption. From power saving perspective, it is better to configure a smaller initial active DL BWP by MIB. However, smaller initial active DL BWP may have lower capacity, e.g., lower paging capacity and lower system information capacity, and thus may impact the system performance. So there is a trade-off between UE power consumption and system performance, when configuring initial active DL BWP.

5.1.2 Paging

Paging reception consumes the main power consumption for NR UEs in RRC idle state and RRC inactive states. For each paging cycle, it may takes the UE up to tens of ms to receive the paging message and whole procedure may include UE's hardware ramp up procedure, re-synchronization procedure and paging PDCCH monitoring and paging PDSCH reception and UE's hardware ramp down procedure. The main power consumption for paging reception includes:

- (Re-)synchronization

Before paging reception, the UE needs to acquire time and frequency synchronization to the gNB. Since there is no "always on" reference signals like LTE CRS available as in LTE, the UE needs to utilize SS/PBCH to get time and frequency synchronization reference. For UEs in bad channels conditions, one-shot time and frequency synchronization using just one SS/PBCH block is not enough, the UE needs to wake up tens of ms before the paging PDCCH monitoring occasion to accumulate several SS/PBCH samples to achieve good enough time and frequency synchronization performance.

The length of paging DRX cycle may have influence on the (Re-)synchronization operation. For example, for short paging cycle, the UE may maintain the synchronization to the network for one or more paging cycles, therefore the UE only needs to perform synchronization every several paging cycles. But for long paging cycle, the UE may need to perform synchronization every paging cycle.

- Paging PDCCH monitoring and paging reception

After (Re-)synchronization to the network, the UE can monitor PDCCH within its paging monitoring occasions. In NR, paging DCI would be often transmitted in beam-sweeping manner. The UE can determine the best PDCCH monitoring occasions to receive the paging DCI for power saving purpose based on the occasion relationships between paging monitoring occasions and SS/PBCH blocks.

For one UE, it would be paged in only very few of paging cycles when there is traffic for the UE or in the case of system information update or ETWS or CMAS, therefore, for most of the paging cycles, paging PDCCH monitoring would result in no valid PDCCH for the UE.

5.1.3 Measurement

RRM measurement is another important power-consuming behaviour for UEs in RRC idle states and RRC inactive states. The UE needs to continuously to perform RRM measurement for cell re-selection.

RRM measurement in RRC idle states and RRC inactive states includes intra-frequency measurement and inter-frequency measurement. Within the SMTC window configured by system information, the UE can perform RRM measurement periodically for each frequency layer. The essential procedures for measurements include：

- Cell detection

For each frequency layer, before performing RRM measurements, the UE needs to detect the target measurement cells. Cell detection would consume much UE's power since the UE needs to search the SS/PBCH block within the SMTC window and at the meantime try to blindly detect the PSS/SSS with at most 1008 hypothetical physical cell IDs.

It is noted that for synchronous network and with the high layer indicated SS/PBCH block measurement locations, the cell detection complexity for the UE in the time domain would be reduced.

- Measurement

The RRM measurement include RSRP measurements and RSRQ measurements. Since multiple-beam deployment would be the typical operation mode for NR, the UE needs to generate the measure results based on the measurements on multiple SS/PBCH blocks which would increase the UE's power consumption.

5.1.4 System information acquisition (including synchronisation)

In LTE, the position and periodicities for PSS, SSS and PBCH are definite and known to a UE. Besides, the SCS for PSS, SSS and PBCH is also definite, i.e., 15KHz. LTE UE may use this information to complete system information acquisition including synchronization. Compared with LTE, there are uncertainties for configurations of SS/PBCH in NR. This may cause complexity and additional power consumption when NR UE performs initial access.

NR UE should support synchronization in time and frequency and the detection of the physical cell IDs from 1008 candidates. Besides, NR UE should support the detection of the SS/ PBCH block under different numerologies and time locations in various frequency carriers and bandwidth configurations. For Sub-6GHz, 15kHz and 30kHz are allowed for SCS of SS/ PBCH block. For above-6GHz, 120kHz and 240kHz are allowed for SCS of SS/ PBCH block. There are several challenging aspects in detecting SS/ PBCH block as mentioned below, and NR UE should be able to handle these challenges.

- In NR, while SS burst set is periodic whose period is assumed to be 20ms by a UE during initial access, there can be multiple SS/ PBCH blocks (up to 8 in sub-6GHz) within each burst set which are not necessarily periodic. As seen in the figure below for 14 (28, respectively) OFDM symbols for 15kHz (30kHz, respectively) SCS, SS/ PBCH block which comprises 4

OFDM symbols is not periodic especially for 30kHz case. In addition to this, not all SS/ PBCH block candidate locations are guaranteed to transmit valid signal.

15kHz SCS	0		1		2		3		4		5		6		7		8		9		10		11		12		13	
30kHz SCS (1st mapping pattern)	0	1	2	3	4	5	6	7	8	9	10	11	12	13	0	1	2	3	4	5	6	7	8	9	10	11	12	13
30kHz SCS (2nd mapping pattern)	0	1	2	3	4	5	6	7	8	9	10	11	12	13	0	1	2	3	4	5	6	7	8	9	10	11	12	13

Figure 5-1 Illustration of SS/PBCH block allocation in time-domain

- 3 or 2 LSB's of SS/ PBCH block within each burst set which a UE needs to identify during initial access is carried by PBCH DMRS scrambling sequence, and this requires the corresponding blind detection for NR UE.
- For some bands, there can be uncertainty on SS/PBCH SCS between 15kHz and 30kHz. UE need perform blind detection.
- Mapping of SS/ PBCH block is described in the figure below. It can be seen that PSS is mapped in the first OFDM symbol, and this creates challenge on AGC operation for PSS detection since there may not be useful signal before PSS which can be used to set proper gain level. However, in LTE, UE can use CRS before PSS.

Figure 5-2 Illustration of SS/PBCH block mapping

- In frequency-domain, unlike LTE whose SS/PBCH is located at the center of system bandwidth, SS/ PBCH block in NR can be flexibly located within each channel. Hence, NR UE would need to be careful if it makes any assumption on spectral shape around SS/ PBCH block.

NR UE should support obtaining the essential minimum system information, including at least SFN, SS block time index and configuration information of PDCCH for RMSI (Remaining minimum System Information) from PBCH.

In NR, detected SS/ PBCH block may not have an associated RMSI, i.e., the SS/PBCH block is not the cell-defining one, and NR UE should be prepared to continue SS/ PBCH block detection at different frequency location with the corresponding indication on PBCH.

5.1.5 RNA update

Compared to LTE, 5G inactive state UEs need to additionally perform RNAU. The size of RAN paging area is a factor contributing to UE power consumption. RAN paging area size can be either the same as or smaller than the Tracking Area. If RAN paging area equals Tracking Area, UE performs combined TAU/RNAU when across TA boundary, no additional signaling/power consumption is expected. However, if RAN paging area is smaller than Tracking Area, more

signaling/power consumption is expected due to extra RNAUs.
Besides, UE needs to perform periodic RNAU. The periodicity of RNAU will also impact UE power consumption. Generally, UE with less activity can be configured with a smaller RNAU periodicity. Also, a decent match of the periodicity of RNAU and TAU can help reduce UE power consumption. Ideally, if they are to have the same length, UE can always perform RNAU and TAU together, power consumption can be minimized. A lesser option is to ensure that the periodicity of TAU is multiple of the periodicity of RNAU, thus TAU can be performed together with RNAU.
UE speed is also a factor that impacts power consumption. The faster the UE speed, more frequent RNAU could be envisioned. One way to counteract this is to make sure one RAN paging area covers UE routes as much as possible by adjusting the RAN paging area at each RNAU based on the history RNAUs. This also applies to adjust the periodicity of RNAU.

5.2 RRC connected state

5.2.1 Bandwidth

The UE channel bandwidth supports a single NR RF carrier in the uplink or downlink at the UE. For Frequency Range 1 (FR1, 450 MHz – 6000 MHz), maximum UE Channel bandwidths for different SCS (sub-carrier spacing) for some typical operating bands in Table 5-1 are supported by NR UE.

Table 5-1 Maximum UE Channel bandwidths per operating band for FR1

FR	Band	15kHz SCS	30kHz SCS	60kHz SCS
FR1	Band n77 (3.3 GHz ~4.2GHz)	50MHz	100MHz	100MHz
	Band n78 (3.3 GHz ~3.8GHz)	50MHz	100MHz	100MHz
	Band n79 (4.4GHz~5GHz)	50MHz	100MHz	100MHz
	Band n1	20MHz	20MHz	20MHz
	Band n3	30MHz	30MHz	30MHz
	Band n8	20MHz	20MHz	NA
	Band n41	50MHz	100MHz	100MHz

For Frequency Range 2 (FR2, 24250 MHz – 52600 MHz), maximum UE Channel bandwidths for different SCS (sub-carrier spacing) for some typical operating bands in Table 5-2 are supported by NR UE.

Table 5-2 Maximum UE Channel bandwidths per operating band for FR2

FR	Band	60kHz SCS	120kHz SCS
FR2	Band n257	200MHz	400MHz
	Band n258	200MHz	400MHz
	Band n260	200MHz	400MHz
	Band n261	200MHz	400MHz

The maximum UE Channel bandwidths for NR FR1 and FR2 is 100MHz and 400MHz, respectively, which is much wider than that of LTE (i.e., 20MHz per carrier). Larger bandwidths will cause more power consumed for NR UE. Not all UEs support the maximum Channel bandwidths. UE can report its supported UL and DL bandwidths to network, by parameters *channelBWs-DL* and *channelBWs-UL,* as shown in Table 5-3, which is narrower than the maximum UE channel bandwidth.

Table 5-3 UE Channel bandwidth capabilities

Definitions for parameters
channelBWs-DL Indicates for each subcarrier spacing whether the UE supports channel bandwidths lower than the maximum channel bandwidth as defined in TS 38.101-1 and TS 38.101-2. If this parameter is not included, the UE supports all channel bandwidths. For FR1, the bits starting from the leading / leftmost bit indicate 5, 10, 15, 20, 25, 30, 40, 50, 60 and 80MHz. For FR2, the bits starting from the leading / leftmost bit indicate 50, 100 and 200MHz.
channelBWs-UL Indicates for each subcarrier spacing whether the UE supports channel bandwidths lower than the maximum channel bandwidth as defined in TS 38.101-1 and TS 38.101-2. If this parameter is not included, the UE supports all channel bandwidths. For FR1, the bits starting from the leading / leftmost bit indicate 5, 10, 15, 20, 25, 30, 40, 50, 60 and 80MHz. For FR2, the bits starting from the leading / leftmost bit indicate 50, 100 and 200MHz.

The BWP operation, by which UE in RRC connected state can operate on a narrower bandwidth than the maximum bandwidth, can also reduce the NR UE power. The BWP concept is descried in Section 5.2.4.

5.2.2 UL-MIMO

UL-MIMO can improve spectrum efficiency compared to single UL transmission. Depending on different deployment scenarios, BS configurations, transmission schemes and SNR distribution etc., an approximately 0%~40% throughput gain could be expected from 2Tx UL MIMO according to simulation.

However, UL-MIMO would require UE to equip multiple RF Tx chains in the uplink. This means multiple sets of PAs, antenna and other components. Compared to single UL, even with the same output power, the power consumption introduced by multiple RF chains would be greater, since activating an extra RF chain would cause more power consumption. According to the power model assumption agreed in 3GPP RAN1 #95 meeting, power consumption of 2Tx is 1.4 times of 1Tx at 0dBm total transmission power, and 1.2 times at 23dBm total transmission power.

The additional power consumption of 2Tx is more significant in lower total transmission power.

In order to reduce the power consumption, fall back to single Tx is a straightforward way, particularly when the scenario is not favorable for 2Tx, i.e. small packet scenario.

5.2.3 HPUE

UE Power Classes define the maximum output power for any transmission bandwidth within the channel bandwidth. Power class 3 of 23dBm, which is reused from LTE, is also the default power class for NR. HPUE (High power UE) means UE supporting higher power class compared to default, e.g. Power class 2 which has a 26dBm maximum output power had been defined for NR bands n41, n77, n78 and n79. Currently, Power class 2 UE was defined for 1Tx (26dBm) and UL-MIMO (2Tx 23+23dBm) SA case. For NSA HPUE, similar definition is being studied.

Table 5-4 NR Band list

NR band	Class 2 (dBm)	Class 3 (dBm)
n41	26	23
n77	26	23
n78	26	23
n79	26	23

The purpose of introducing HPUE feature is to improve the uplink coverage. There is an imbalance between uplink and downlink coverage and generally UL coverage is weaker than DL, and this became more apparent for 5G, since the spectrum available for 5G is usually higher compared to LTE. As an example, the following NR link budget result was referenced from [R1-1809266 IMT-2020 Self-Evaluation: NR Link Budgets] in which a lower coupling loss means a weaker coverage. It can be seen that the coupling loss gap between UL and DL is relatively high especially for 3.5GHz.

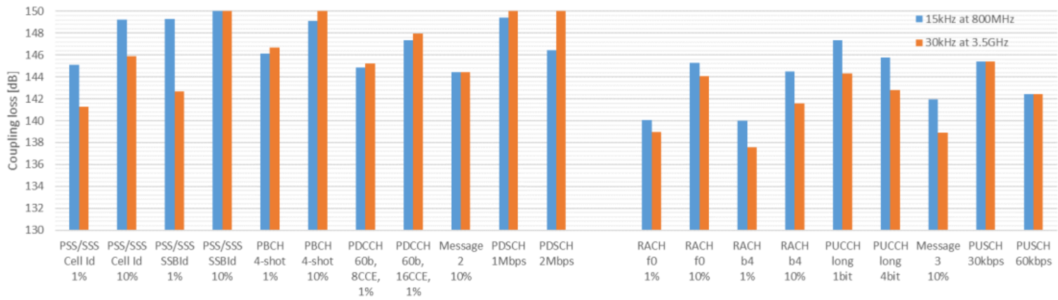


Figure 5-3 Preliminary results of maximum coupling loss for 15kHz numerology at 800MHz and 30kHz numerology at 3500MHz.

It should be noted that, in order to control emission to satisfy SAR requirements, only TDD band has the definition for HPUE and there is also a restriction in UL duty cycle to control the actual transmission time.

In the case of identical uplink path-loss, a PC2 HPUE higher peak power delivers higher uplink SNR, which may allow uplink scheduler to allocate higher MCS, which may translate into shorter transmission time, and in turns may reduce power consumption especially for large uplink packet transmissions. This effect is more notice-able at the PC3 UE cell edge location and for large uplink data payloads. As the PC3 UE reaches cell edge, the lower peak power of a PC3 UE forces the uplink scheduler to decrease DTX rate, i.e. the PC3 UE transmission time might be longer than that of a PC2 UE (assuming same uplink path loss) resulting into a higher battery power consumption. Nevertheless, the relationship between peak power capability and UE battery life is not straightforward as it depends on many factors, such as UE quality of service, packet size, scheduler algorithm, cell loading, etc… etc. Considering the large number of variable, further studies would be needed to evaluate user experience for a variety of use-cases. Alternatively, the higher peak power capability of PC2 could be used by operators as a way to enhance cell coverage, making this feature very attractive. However the extended coverage may come at the expense of a higher battery power consumption.

5.2.4 BWP

The Bandwidth Part (BWP) controls the size and location of the bandwidth used by the DL/UL links. The size of the BWPs used for DL/UL transmissions are configured independently for the DL/UL. The UL bandwidth size has a relatively small impact on the power consumption of the UE which is mainly controlled by the transmit power (see [4]). On the other hand, the bandwidth used for DL reception has a large impact on the power consumption as it controls the RF, the ADC sampling rate, the FFT size and the baseband processing as well as the peak throughput. Based on the NR UE power model recently agreed in 3GPP [3], lowering the bandwidth from 100MHz to 20MHz reduces UE downlink power consumption by ~60% (see Figure 5-4).

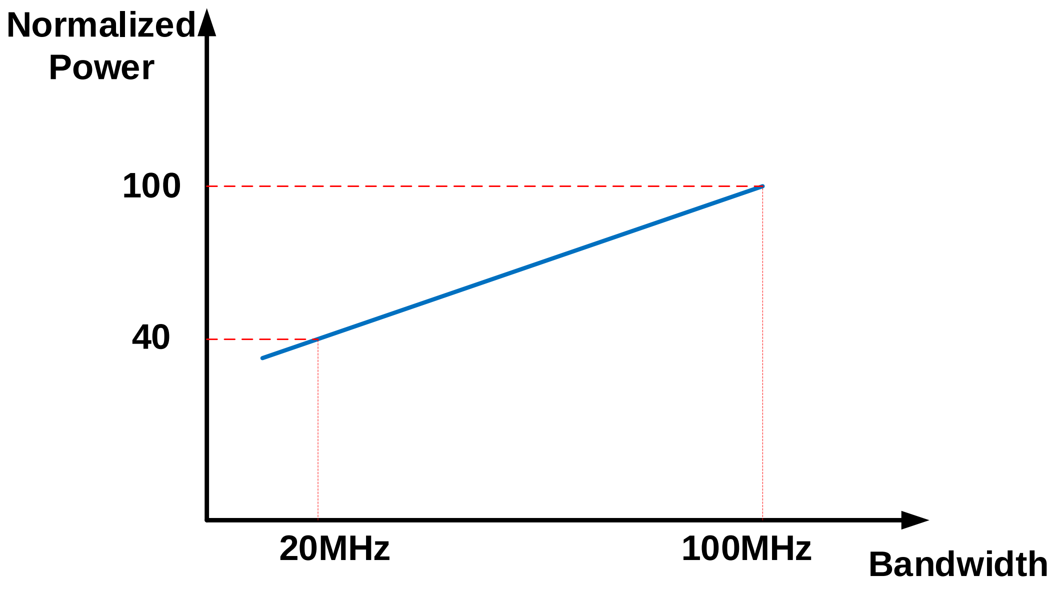


Figure 5-4 UE power consumption scaling with the DL bandwidth for 4Rx MIMO reception

In addition to the BWP size, the BWP configuration can also specify several other parameters with impact on the power consumption (check Table 5-5 for summary):

- Cross-slot scheduling: DL cross-slot scheduling is enabled or disabled according to the k0 values specified in the TimeDomainResourceAllocationList in PDSCH-ConfigCommon in BWP-DownlinkCommon [11]. The power consumption saving from cross-slot scheduling is estimated at ~30% (see [3]), however this can be only achieved if the UE is guaranteed that same slot scheduling cannot be used by the gNodeB which requires (K0≥1) for the applicable PDSCH time domain resource allocation. To allow an effective use of cross-slot scheduling in Rel-16, a change of the framework to support an explicit signaling of the minimum K0 per power profile (Figure 5-5) will allow more flexibility and a more direct mechanism to control the power consumption

- Number of MIMO layers: rank restriction for CSI reporting per BWP configuration are possible in Rel-15. This allows the UE to reduce the CSI reporting complexity and can be used to limit the maximum number of MIMO layers that the UE has to report and expect to receive. However, in Rel-15 this mechanism does not prevent the gNodeB from using a higher number of MIMO layers or requiring performance with the maximum number of Rx antennas supported by the UE, and this limits the power savings that the UE can achieve. Therefore to improve power saving, further mechanisms of adaptation of the maximum number of MIMO layers and number of Rx antennas should be supported in Rel-16 as part of the power profiles (Figure 5-5).
- PDCCH monitoring: PDCCH monitoring is controlled through the DL BWP configuration. When no data is transmitted (PDCCH-only), the power consumption is dominated by the PDCCH processing, hence reducing PDCCH monitoring improves the power consumption significantly. The main mechanism for PDCCH monitoring reduction is DRX (see section 5.2.5), but switching between BWPs with different bandwidths and PDCCH monitoring periodicities also enables adaptation of the UE configuration to the traffic and saving power when latency requirements are relaxed.

Table 5-5 Main parameters controlled by the BWP with impact on power consumption.

Parameters controlled by the BWP	Impact on power consumption
Bandwidth size	UE power consumptions scales with the DL bandwidth (see Figure 5-4)
Cross-slot scheduling	Cross slot scheduling (K0≥1) reduces power consumption (see section 5.2.4)
Number of MIMO layers	In Rel-15, partial control of the number of MIMO layers through the CSI configuration per BWP Controlled by the DL BWP
PDCCH monitoring	Adaptation of the PDCCH monitoring to the traffic allow to reduce power consumption.

Figure 5-5 shows an example of how the BWP can be used to build different power profiles that can then be used to adapt the UE configuration to the traffic characteristics, therefore achieving savings on power consumption without impact the quality of service. Some of building blocks for realizing power profiles similar to those of Figure 5-5 are already present in Rel-15 BWP configurations, however, improvements to the framework such as minimum K0 and maximum MIMO configurations (#layers and ~Rx antennas) are needed in Rel-16 to improve effectiveness.

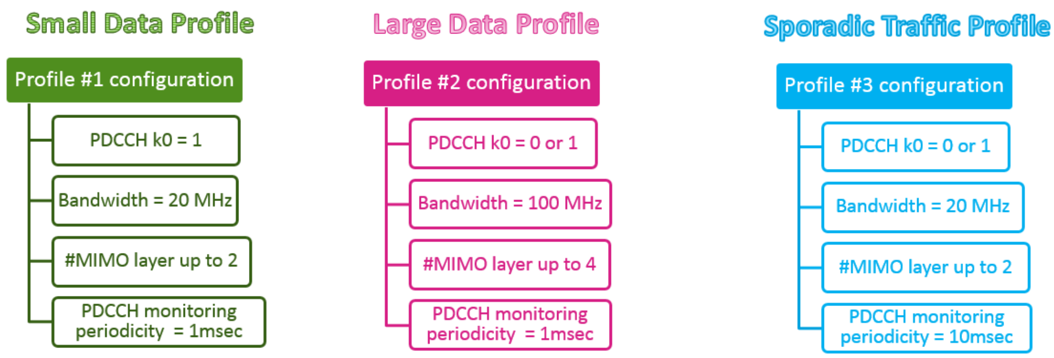


Figure 5-5 Examples of power profiles for different traffic characteristics

Rel-15 supports up to 4 BWPs for each of the DL/UL and the switching between these BWPs is either RRC-based or DCI-based. For Rel-16, the switching between power profiles (Figure 5-5) can be based on the same or similar mechanisms to the BWP switching, however given that the RRC-based switching is typically order of magnitude slower than DCI-based switching (tens of ms compared to few ms), therefore to achieve fast adaptation and lower traffic disruption when switching between the power profiles the support of a DCI-based switching is essential.

5.2.5 DRX

DRX is one scheme to save the UE's power consumption for UE in RRC connected state. The power consumption factors for DRX procedure are as follows:

- (Re-)synchronization procedure

For DRX, time and frequency (Re-)synchronization as discussed in the paging section 5.1.2 is a key power consumption procedure for DRX. When the DRX cycle is long, the UE needs to utilize SS/PBCH blocks to acquire time and frequency (Re-)synchronization to the network.

- DRX parameters setting

Proper DRX parameter setting is very important for UE's power consumption. It would be beneficial to configure the DRX cycle, the length of on-duration and DRX related timers to match the UE's traffic characteristics to save the UE's power.

Discontinuous Reception (DRX) design for Rel-15 NR, is very similar to E-UTRA DRX. Using DRX enables power saving by reducing the time UE is awake to monitor the PDCCH.
The UE is only required to monitor the PDCCH during the DRX-OnDuration. If the UE detects a DCI with a DL assignment or an UL grant, then the UE restarts the DRX-inactivity timer during which UE keeps monitoring the PDCCH, otherwise the UE may stop monitoring the PDCCH at the end of the DRX-OnDuration and go back to sleep.

The power reduction achieved is proportional to the time the UE spends in sleep mode, and so the OnDuration and Inactivity timers should be short in relation to the DRX cycle time. TTI duration is inversely proportional to subcarrier spacing, so for a given timer setting, an NR UE will process 2^{μ} times as many TTIs as its E-UTRA equivalent. This means that unless traffic patterns dictate otherwise, shorter timer settings should be the norm in NR. If this cannot be achieved, additional signalling can be used to provide early termination of the DRX cycle when possible.

In addition to the PDCCH monitoring, in order for the UE to maintain connection and for the network to optimize resources utilization, the network needs to configure a number of reference signals and control channel transmissions, while the UE needs to perform the corresponding Background Activity (BA) tasks in connected DRX.
The background tasks that the UE has to perform in C-DRX can be divided into 2 groups:

- Background Activity during the OnDuration:
    - This group of background activities includes CSI-RS acquisition, Beam Management, CSI reporting and SRS transmission tasks. These tasks can only take place during the OnDuration.
- Background Activity outside of the OnDuration:
    - This group of background activities includes the SSB acquisition/processing, TRS acquisition/processing and RRM measurement. While the SSB/TRS processing has to take place before the OnDuration for the UE to be ready to start receiving the PDCCH/PDSCH, the RRM measurement timing depends on the SMTC configuration, although the UE maintains some flexibility on which SMTC occasion to use.

The setting of Figure 5-6 shows that the power consumed by the BA tasks outside of the OnDuration is of the same order of magnitude as the power consumed during the OnDuration. The BA tasks during the OnDuration lead to a relatively small power consumption increase, but still occupy a significant portion of the OnDuration hence putting a limit on how small the OnDuration could be made.

To achieve better power consumption, optimization of the BA tasks should be targetted in NR Rel-16. Power saving could be realized by grouping the CSI acquisition/reporting and potentially other BA activities with SSB/TRS processing in a pre-wakeup window. Such a mechanism would allow reduction of the DRX-OnDuration length, which would improve performance and save power.

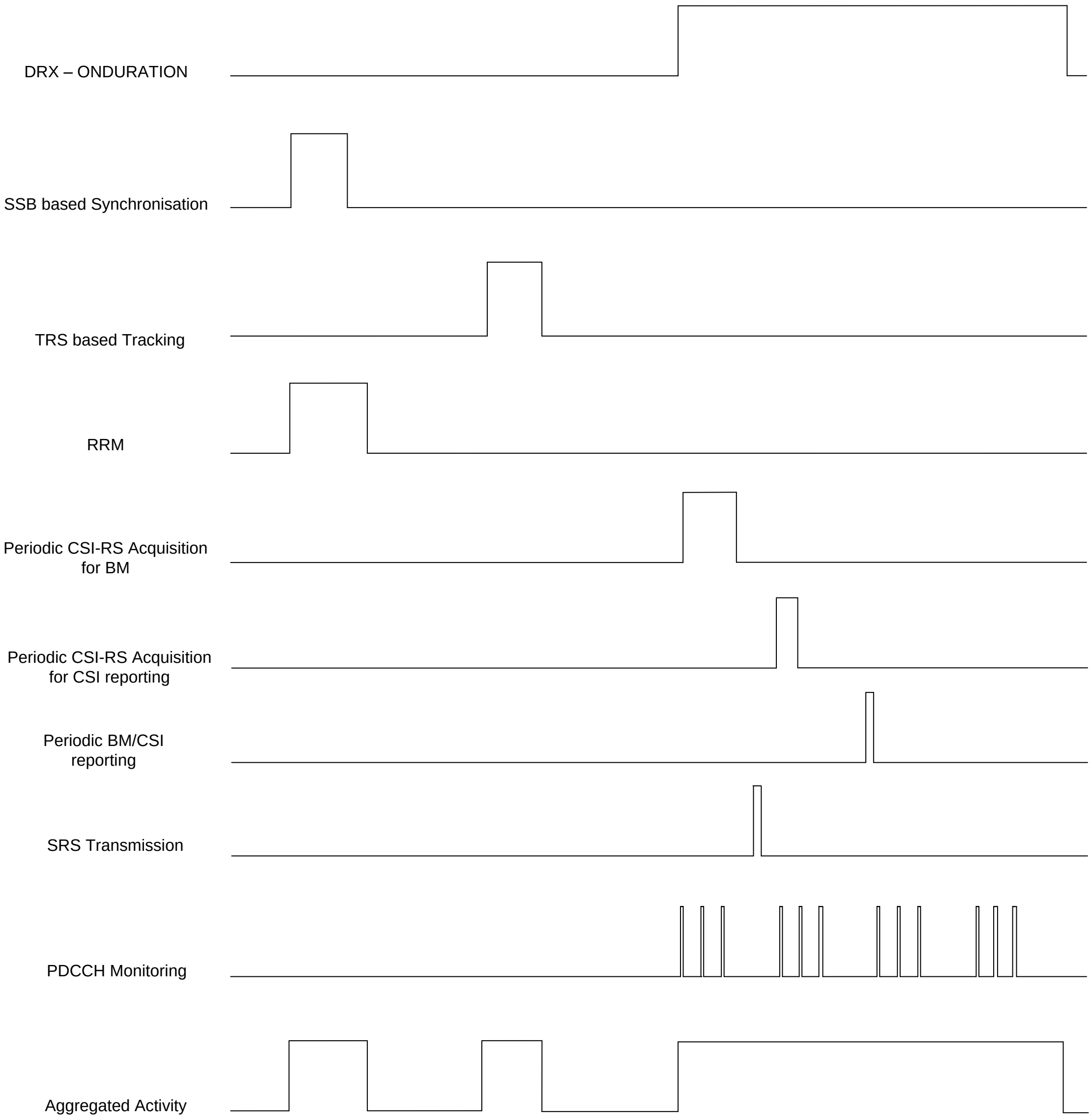


Figure 5-6 Processing Timeline including the Background activity.

5.2.6 Cross-slot scheduling

NR support flexible timings between PDCCH and PDSCH and the timing K0 is informed in the DCI from one RRC configured set named PDSCH-TimeDomainResourceAllocationList. If the minimum value of k0 within the PDSCH-TimeDomainResourceAllocationList is 0, the UE can be scheduled with same-slot PDSCH scheduling by indicating that k0 equals to 0 or with cross-slot scheduling by indicating that k0 is larger than 0 from that set. In this case, the UE needs to receive and buffer all the PRBs in the current active BWP in order to be ready to receive PDCCH and PDSCH within the same slot. Otherwise, if the minimum value of k0 within the PDSCH-TimeDomainResourceAllocationList is larger than one K_threshold, only cross-slot scheduling would be allocated for the UE. In this case, the UE can tune its RF with a narrow

bandwidth that is wide enough to receive and buffer the PDCCH CORESET, then if one DCI is detected, the UE can tune its RF to the cover the bandwidth of PDSCH resource. By such operation, the UE can save its power when there is no PDSCH scheduling. It is noted that K_threshold shall be large enough to guarantee that the UE has enough time to switch its RF.
Therefore, for UE's services that are not latency sensitive, cross-slot scheduling is beneficial for UE's power saving. And the key point to achieve the power saving target is that configure that PDSCH-TimeDomainResourceAllocationList with all the elements in the list are larger than one K_threshold.

5.2.7 Periodic PDCCH monitoring

The most power consuming procedure for one NR UE would be PDCCH monitoring in RRC connected state, as reported by several chipset vendors, for LTE, more than half of the PDCCH monitoring occasion corresponds to no PDSCH schedule. NR support much more flexible PDCCH search space sets configuration to match the UE's traffic characteristic. However, the UE's traffic may fluctuates time by time, it is hardly possible for the gNB to use only one set of PDCCH search space configuration for one UE all the time. Furthermore, the network often needs to schedule many UEs at the same time thus it cannot be guaranteed that one UE could be scheduled when the system load is high. Therefore, reducing power waste for PDCCH monitoring is one important direction for NR UE' power saving. There are several factors affect the power consumption for PDCCH monitoring.

- Match to the UE's traffic characteristic

The more the PDCCH search space configuration matches the UE's traffic characteristic, the less the power waste. Therefore, it is important for the network to have a clear knowledge of the UE's traffic characteristic for different services.

- Fast switching of PDCCH search space configuration

In order to adapt to the fluctuation of the UE's traffic, it is important to switch the UE's PDCCH search space configuration as fast as possible to reduce the chances of PDCCH monitoring without corresponding PDSCH scheduling. The current PDCCH search space reconfiguration mechanism based on RRC reconfiguration may not meet such requirements and needs to be enhanced.

5.3 Parameter configuration recommendation

In this section we consider how parameter configuration affects UE power consumption in different scenarios, and show using examples that the best configuration to minimise UE power consumption depends on the type of data traffic that the UE is experiencing. The evaluation is based on work done in support of the 3GPP study item on NR UE power consumption.

In this study, consensus was reached on a UE power model that can be used to compare the relative benefits of different NR power saving strategies. Full details of the model can be found in [13], but the key features are summarized below.

5.3.1 3GPP power model

The model reference configuration for FR1 is based on a single 100MHz TDD carrier with 30kHz SCS. In the downlink the UE has 4 Rx antennas. PDCCH occupies 2 symbols at the start of a slot, using same-slot scheduling (k0=0), and PDSCH operates at maximum data rate (256QAM). The UE uplink has 1 Tx antenna transmitting at 0dBm or 23dBm.

Power values are averaged over 1 slot, and the model specifies the average power consumption in each of several power states, listed in Table 5-6. The model is based on an uncalibrated power unit representing average UE power in the deep sleep state, which makes it independent of any specific UE implementation. Although different UE implementations are expected to depart from the model ratios for the relative power consumption of each state, simulations using the model provide a platform-independent reference point for evaluation and comparison of different power saving strategies.

Power State	Characteristics	Relative Power
Deep Sleep	Time interval for the sleep should be larger than the total transition time entering and leaving this state. Accurate timing may not be maintained.	1 (Optional: 0.5)
Light Sleep	Time interval for the sleep should be larger than the total transition time entering and leaving this state.	20
Micro sleep	Immediate transition is assumed for power saving study purpose from or to a non-sleep state	45
PDCCH-only	No PDSCH and same-slot scheduling; this includes time for PDCCH decoding and any micro-sleep within the slot.	100
SSB or CSI-RS processing	SSB can be used for fine time-frequency sync. and RSRP measurement of the serving/camping cell. TRS is the considered CSI-RS for sync. FFS the power scaling for processing other configurations of CSI-RS.	100
PDCCH + PDSCH	PDCCH + PDSCH. ACK/NACK in long PUCCH is modeled by UL power state.	300
UL	Long PUCCH or PUSCH.	250 (0 dBm) 700 (23 dBm)

Table 5-6 3GPP model power states

The model also includes transitional energies for the ramp down + ramp up energy associated with entering and leaving a sleep state, listed in Table 5-7.

Sleep type	Additional transition energy: (Relative power x ms)	Total transition time
Deep sleep	450	20 ms
Light sleep	100	6 ms
Micro sleep	0	0 ms*
* Immediate transition is assumed for power saving study purpose from or to a non-sleep state		

Table 5-7 3GPP model transition energies

By modelling the time spent in each state and the number of sleep transitions in a given period, using system level simulation, the average power consumption for the reference configuration can be calculated. When the reference configuration is modified by enabling cross-slot scheduling or by changing BWP, the number of carriers or the number of antennas, the model provides a set of scaling factors to be applied to the Table 5-6 parameters. These are listed in Table 5-8.

Scaling for FR1	Proposal	Comment
BWP Bandwidth (DL)	Scaling of X MHz = 0.4 + 0.6 * (X - 20) / 80. Linear interpolation for intermediate bandwidths. Valid only for X = 10, 20, 40, 80, and 100. Above scaling is applicable for FR1 only. In case scaling is needed for FR2, companies can report the assumed scaling factor.	For 10MHz BW, only AL up to 8 can be used for PDCCH The transition time is the same as DCI-based BWP switching delay for Rel-15. If the power after scaling is smaller than the BWP transition power, assume the BWP transition power as the output of scaling unless otherwise justified.
BWP Bandwidth (UL)	No scaling at 0dBm or 23dBm Above scaling is applicable for FR1 only. In case scaling is needed for FR2, companies can report the assumed scaling factor.	
CA (DL)	2CC is 1.7x1CC 4CC is 3.4x1CC (i.e. 2x 2CC) Above refers to the worst case CA configuration in terms of power consumption.	Activation/deactivation delay follows RAN4 specification; FFS transition energy Applicable for FR1 and FR2
CA (UL)	Same as downlink at 0dBm. No scaling at 23dBm 2CC is 1.2x1CC at 23dBm Limit scaling up to 2CC.	Applicable for FR1 and FR2
Antenna scaling (DL)	2Rx power is 0.7x 4Rx power for FR1 1Rx power is 0.7x 2Rx power for FR2	Assume same number of antenna elements per Rx chain
Antenna scaling (UL)	2Tx power is 1.4x 1Tx power at 0dBm. 1.2x.at 23dBm FR1 only	2Tx support is not considered for FR2.
PDCCH-only	Power of cross-slot scheduling is 0.7x same-slot scheduling	Applicable for FR1 and FR2
SSB	One SSB power is 0.75 of two SSB power, i.e. 75 power units	
PDSCH-only slot	280 for FR1 325 for FR2	This assumes the same number of PDSCH symbols as in the PDCCH+PDSCH case.
CSI-RS	FFS for scaling w.r.t. # of symbols for CSI-RS	
Short PUCCH	Short PUCCH power = 0.3 x uplink power Reference config consists of 1-symbol PUCCH	Applicable for FR1 and FR2.
SRS	SRS power = 0.3 x uplink power	Applicable for FR1 and FR2.

Table 5-8 3GPP model scaling parameters

It should be noted that for BWP scaling in the downlink, the final comment modifies the proposal. This imposes a minimum value of 50 units on the UE power after BWP reduction is applied, which affects PDCCH-only and SSB/CSI-RS power when the BWP is 20MHz or less.

5.3.2 Applying the 3GPP power model to NSA

In its currently agreed form the 3GPP power model is only applicable to standalone configurations. Most initial 5G deployments will be non-standalone, and in this paper we also want to discuss power consumption for these configurations. It is therefore necessary to make some assumptions about the power model that will be applied to the NSA configuration.

For the analysis presented here, we have used the following assumptions to align the LTE component modelling with the 3GPP NR model, while avoiding adding new power states to the model.

- The LTE modem and the NR modem operate independently in the UE (they may in fact be separate devices)
- Sleep states and transition energies use the same parameters in both modems
- The LTE modem in active states uses the same parameters as an NR modem operating in a 20MHz BWP.
- LTE sync and measurements use the same power as NR SSB/CSI-RS for sync/reference symbol processing
- When LTE combines PDCCH and uplink transmission in the same slot, the power is the sum of the two component powers
- Because the LTE modem only has 2 Rx paths instead of 4, the 50 unit floor is not applied to the BWP scaling for PDCCH-only and Sync/RS monitoring

5.3.3 Modelling scenarios

The analysis in [13] is based around 3 reference traffic scenarios (Table 5-9). Results are presented from many companies, firstly using the baseline model configuration and then applying different power saving configurations.

	FTP traffic	Instant messaging	VoIP
Model	FTP model 3	FTP model 3	As defined in R1-070674. Assume max two packets bundled.
Packet size	0.5 Mbytes	0.1 Mbytes	
Mean inter-arrival time	200 ms	2 sec	
DRX setting	Period = 160 ms Inactivity timer = 100 ms On duration = 8 ms	Period = 320 ms Inactivity timer = 80 ms On duration = 10 ms	Period = 40 ms Inactivity timer = 10 ms On duration = 4 ms

Table 5-9 Traffic scenarios for modelling

NR allows considerable flexibility in configuration, and the savings reported vary widely between companies, but all of the results indicate that by adapting the network configuration to UE traffic patterns, significant savings can be achieved.

In this paper the same reference scenarios are used as the baseline for power comparisons. Standalone results are based on the simulation timings in [14], using a simplified power calculation to examine the savings available from DRX cycle configuration, BWP adaptation, cross-slot scheduling and MIMO reduction.

Power saving techniques applied to standalone cases would give reductions of similar magnitude if applied to the same traffic on the NR carrier of an NSA network, but the percentage saving would be smaller in the NSA case due to the added power contribution of the LTE component.

For an NSA network the configuration flexibility is even greater than for SA, and to avoid a multiplicity of scenarios we have considered here an example of mixed IM and FTP traffic, with IM traffic (mostly small packets) routed over the 20MHz LTE carrier, and the FTP traffic (mostly large packets) routed over the 100MHz NR carrier. This is compared with an SA configuration where both sets of traffic are routed over a single 100MHz NR carrier.

The results from system level simulation are presented in subsequent sections.

5.3.4 Standalone baseline scenarios

The first of the reference scenarios is FTP traffic, which is representative of applications such as file download or streaming video. The simulation output is summarised in Table 5-10 below.
The simulation averages results from multiple UEs in a cell, so the number of sleep transition instances in a 10 second period can be fractional.

FTP baseline (0.5MB packets, 200ms arrival, DRX 160, 100, 8)				
Power state	Model power	Time percentage	Average power	% power
Microsleep	45	0.48%	0.22	0.42%
Light sleep	20	5.17%	1.03	1.99%
Deep sleep	1	55.65%	0.56	1.07%
PDCCH-only	100	33.13%	33.13	63.73%
PDCCH+PDSCH	300	4.09%	12.27	23.60%
SSB/CSI-RS	100	1.31%	1.31	2.52%
UL	250	0.16%	0.40	0.77%
Transitions	Model energy	Instances/10sec	Average power	% power
LS transitions	100	38.99	0.39	0.75%
DS transitions	450	59.55	2.68	5.15%
		Total Power	**51.99**	**100%**

Table 5-10 FTP traffic model baseline power consumption

The most power-hungry states are not necessarily the biggest contributors to average power consumption. 63.7% of average power consumption results from monitoring PDCCH waiting for new packets to arrive. Downlink data traffic (PDCCH+PDSCH) has the highest power consumption at 300 units, but because it accounts for just over 4% of the time allocation it only contributes 24% of total power consumption. Uplink traffic power is slightly lower (250 units) and accounts for only 0.8% of total power consumption, but even at 23dBm (power level 700 units) this would only rise to 2.1%. Sleep states, although they occupy over 60% of the time allocation, account for only 3.5% of average power consumption, but the sleep transitions contribute an additional 5.9%.

This pattern is typical of many representative use cases. The highest throughput cases that result in UE overheating will be relatively infrequent for most users.

In this scenario a large proportion of the PDCCH monitoring power is spent waiting for the DRX inactivity timer to expire. It will be shown later that the DRX cycle is suboptimal for this type of traffic, and by adapting the DRX cycle to the traffic patterns significant reductions in power consumption can be achieved.

The next reference scenario is Instant messaging traffic, which is characterised by a smaller packet size and a longer mean delay between successive packet arrivals. The simulation results are shown in Table 5-11.

IM baseline (0.1MB packets, 2000ms arrival, DRX 320, 80, 10)				
Power state	Model power	Time percentage	Average power	% power
Microsleep	45	0.23%	0.10	0.96%
Light sleep	20	2.88%	0.58	5.34%
Deep sleep	1	89.79%	0.90	8.32%
PDCCH-only	100	6.26%	6.26	58.04%
PDCCH+PDSCH	300	0.06%	0.18	1.67%
SSB/CSI-RS	100	0.70%	0.70	6.49%
UL	250	0.08%	0.20	1.85%
Transitions	Model energy	Instances/10sec	Average power	% power
LS transitions	100	21.34	0.21	1.98%
DS transitions	450	36.78	1.66	15.35%
		Total Power	**10.79**	**100%**

Table 5-11 Instant messaging traffic model baseline power consumption

The reduced data packet frequency means that power consumption in this case is significantly lower – just over 20% of the FTP case. PDCCH monitoring is still the biggest contributor to the average, accounting for 58% of the total, and sleep/sleep transitions account for a further 32%. Of the remaining 10%, 6.5% comes from synchronisation and measurement activities, with only 3.5% resulting from user data transfer.

The power from sleep transitions is again higher than the power in sleep states - reducing the number of wakeup occasions (for example by lengthening the DRX cycle, or by grouping activities closer together in time) would make a significant reduction in the sleep contribution.

The last reference scenario is voice traffic, characterised by very small data packets closely spaced in time, requiring frequent wakeups to satisfy latency requirements. Results for this case are shown in Table 5-12 below.

VoIP baseline (R1-070674, DRX 40, 10, 4)				
Power state	Model power	Time percentage	Average power	% power
Microsleep	45	2.99%	1.35	2.68%
Light sleep	20	43.53%	8.71	17.32%
Deep sleep	1	24.11%	0.24	0.48%
PDCCH-only	100	22.85%	22.85	45.45%
PDCCH+PDSCH	300	1.25%	3.75	7.46%
SSB/CSI-RS	100	5.10%	5.10	10.14%
UL	250	0.16%	0.40	0.80%
Transitions	Model energy	Instances/10sec	Average power	% power
LS transitions	100	322.26	3.22	6.41%
DS transitions	450	103.51	4.66	9.27%
		Total Power	**50.27**	**100%**

Table 5-12 VoIP traffic model baseline power consumption

Although the average data rate is only kilobits per second, power consumption in this case is almost as high as for FTP traffic. PDCCH-only power is still the largest contributor at 45%, but the short DRX cycle means that the number of wakeups increases dramatically, and the short sleep duration prevents deep sleep in many cases. Sleep and sleep transitions account for just over 36% of the total.

When the traffic data rate is very low it is wasteful for the UE to process the entire 100MHz bandwidth of the carrier. Savings can be made by using a narrower BWP and/or fewer active antennas to reduce the processing load, or by reducing the duration of the active portion of the DRX cycle.

The effectiveness of different power saving techniques is examined in the next section

5.3.5 Standalone power saving evaluation

In each of the examples that follow, a single feature of one of the baseline scenarios is changed, and the resulting power saving is evaluated. By combining multiple power saving techniques in a single configuration, greater savings are possible, but the 3GPP model is likely to overestimate the combined saving in such cases since some savings may be duplicated in different scaling factors – if two techniques each deliver a power saving of 40% when applied individually, the saving when both are combined may be less than the expected 64%.

5.3.5.1 DRX cycle adaptation

The DRX cycle applied to FTP traffic in the baseline case is not well matched to the data patterns. The cycle duration is close to the mean packet arrival time, so a large proportion of cycles will contain data, triggering the inactivity timer and forcing the UE to be active for 100ms of the 160ms cycle. Reducing the inactivity timer duration will increase the proportion of time that the

UE can spend asleep, and reduce the time spent monitoring the control channel. Table 5-13 shows the result of reducing the inactivity timer from 100ms to 20ms.

FTP with DRX adaptation (DRX 160, 20, 8)				
Power state	Model power	Time percentage	Average power	% power
Microsleep	45	0.52%	0.23	0.78%
Light sleep	20	5.45%	1.09	3.62%
Deep sleep	1	78.10%	0.78	2.60%
PDCCH-only	100	10.32%	10.32	34.30%
PDCCH+PDSCH	300	4.04%	12.12	40.28%
SSB/CSI-RS	100	1.41%	1.41	4.69%
UL	250	0.16%	0.40	1.33%
Transitions	Model energy	Instances/10sec	Average power	% power
LS transitions	100	41.16	0.41	1.37%
DS transitions	450	73.79	3.32	11.04%
Reduction	**42.13%**	**Total Power**	**30.09**	**100%**

Table 5-13 Reducing FTP traffic power consumption with DRX adaptation

This single change reduces average power consumption by 42%. The power required for uplink and downlink user data transfer (12.5units) has not changed significantly, but the time that the UE spends asleep has increased from 61% to 84%, and the power due to PDCCH-only monitoring has fallen from 33.1 units to 10.32 units so that it is no longer the most significant user of power. The only performance penalty is a small increase in latency for those data packets that arrive while the UE is asleep.

More aggressive reductions in the inactivity and on-duration timer settings would result in further savings.

For instant messaging traffic, the DRX cycle is significantly shorter than the mean interval between consecutive packets. If the application can tolerate increased latency, a longer DRX cycle will save power. A longer DRX cycle increases the probability that data will arrive during the DRX cycle to trigger the inactivity timer, so reducing the inactivity timer setting will provide further savings. Table 5-14 shows the result of increasing the DRX cycle from 320ms to 1280ms, and at the same time reducing the inactivity timer from 80ms to 20ms.

IM with DRX adaptation (DRX 1280, 20, 10)				
Power state	Model power	Time percentage	Average power	% power
Microsleep	45	0.06%	0.03	0.84%
Light sleep	20	0.72%	0.14	4.46%
Deep sleep	1	97.75%	0.98	30.25%
PDCCH-only	100	1.22%	1.22	37.75%
PDCCH+PDSCH	300	0.05%	0.15	4.64%
SSB/CSI-RS	100	0.18%	0.18	5.57%
UL	250	0.02%	0.05	1.55%
Transitions	Model energy	Instances/10sec	Average power	% power
LS transitions	100	5.32	0.05	1.65%
DS transitions	450	9.56	0.43	13.31%
Reduction	**70.04%**	**Total Power**	**3.23**	**100%**

Table 5-14 Reducing IM traffic power consumption with DRX adaptation

This change reduces instant messaging power consumption by 70%. The time that the UE is asleep increases from 93% to 98.5%, and PDCCH-only power consumption reduced from 6.3 units to 1.2 units.

These two examples show that large reductions in UE power consumption can be obtained by adapting DRX parameters to traffic patterns. The power saving is usually accompanied by an increase in latency, so the longest acceptable DRX cycle is usually constrained by application performance requirements. Once the DRX cycle has been set, if there is a high probability that data will arrive in one DRX period then reductions in the inactivity timer setting will be most effective in reducing power. If the probability is low, then reductions in the on-duration timer setting will be more effective.

The short DRX cycle required for VoIP traffic means that there are fewer opportunities for significant savings from DRX adaptation alone. For this case, semipersistent scheduling can reduce signaling traffic, and bandwidth and MIMO reductions will reduce the processing load. Locating measurement resources near to the start of the DRX period will minimize the duration of the active portion of the DRX cycle, and this may allow more efficient use of sleep states. After these options are in place, DRX adaptation may offer further savings.

5.3.5.2 Cross-slot scheduling

If PDCCH can schedule PDSCH resources in the same slot that it is received (same-slot scheduling) , then downlink symbol reception and capture must continue until PDCCH decoding is complete, in case the captured symbols contain data for the UE. If the decoded PDCCH does not schedule data for the UE in that slot, then the energy spent doing this has been wasted.

In NR, PDCCH can schedule PDSCH resources in the next slot, or a later slot, by setting parameter k0 to a value greater than zero (cross-slot scheduling). If cross-slot scheduling is enabled, the UE can disable its receive path as soon as the last sample of the last symbol of PDCCH has been received, and when PDCCH decoding is complete it can enter microsleep until the start of the next slot. In the 3GPP model this reduces power consumption in PDCCH-only slots by 30%.

Simulations with cross-slot scheduling enabled generate the same set of time percentages as the baseline scenarios, but PDCCH-only power consumption is reduced from 100 units to 70 units. Since PDCCH-only power is the largest component in each of the baseline scenarios, this produces significant savings. Table 5-15 - Table 5-17 show the results from each of the three reference scenarios.

FTP traffic with Cross-slot scheduling				
Power state	Model power	Time percentage	Average power	% power
Microsleep	45	0.48%	0.22	0.51%
Light sleep	20	5.17%	1.03	2.46%
Deep sleep	1	55.65%	0.56	1.32%
PDCCH-only	70	33.13%	23.19	55.15%
PDCCH+PDSCH	300	4.09%	12.27	29.18%
SSB/CSI-RS	100	1.31%	1.31	3.12%
UL (0dBm)	250	0.16%	0.40	0.95%
Transitions	Model energy	Instances/10sec	Average power	% power
LS transitions	100	38.99	0.39	0.93%
DS transitions	450	59.55	2.68	6.37%
Saving	**19.12%**	**Total power**	**42.05**	**100.00%**

Table 5-15 Reducing FTP traffic power consumption with cross-slot scheduling

IM traffic with Cross-slot scheduling				
Power state	Model power	Time percentage	Average power	% power
Microsleep	45	0.23%	0.10	1.16%
Light sleep	20	2.88%	0.58	6.47%
Deep sleep	1	89.79%	0.90	10.08%
PDCCH-only	70	6.26%	4.38	49.19%
PDCCH+PDSCH	300	0.06%	0.18	2.02%
SSB/CSI-RS	100	0.70%	0.70	7.86%
UL (0dBm)	250	0.08%	0.20	2.25%
Transitions	Model energy	Instances/10sec	Average power	% power
LS transitions	100	21.34	0.21	2.40%
DS transitions	450	36.78	1.66	18.58%
Saving	**17.41%**	**Total power**	**8.91**	**100.00%**

Table 5-16 Reducing IM traffic power consumption with cross-slot scheduling

VoIP traffic with Cross-slot scheduling				
Power state	Model power	Time percentage	Average power	% power
Microsleep	45	2.99%	1.35	3.10%
Light sleep	20	43.53%	8.71	20.05%
Deep sleep	1	24.11%	0.24	0.56%
PDCCH-only	70	22.85%	16.00	36.84%
PDCCH+PDSCH	300	1.25%	3.75	8.64%
SSB/CSI-RS	100	5.10%	5.10	11.75%
UL (0dBm)	250	0.16%	0.40	0.92%
Transitions	Model energy	Instances/10sec	Average power	% power
LS transitions	100	322.26	3.22	7.42%
DS transitions	450	103.51	4.66	10.73%
Saving	**13.64%**	**Total power**	**43.42**	**100.00%**

Table 5-17 Reducing VoIP traffic power consumption with cross-slot scheduling

The 30% reduction in PDCCH monitoring power gives power reductions of between 13% and 19% for these reference scenarios. The 3GPP model calibration results in appendix 8.2 of [13] show clearly that if C-DRX is not enabled, the time percentage for PDCCH-only slots is close to 100%, so for non- DRX scenarios cases the cross-slot savings could approach 30%.

5.3.5.3 Bandwidth part adaptation

The processing requirement in NR UEs is related to the number of subcarriers that have to be processed, and therefore to the BWP bandwidth. A smaller bandwidth means a lower sampling rate and a smaller FFT, and this has a direct relationship to power consumption. However, it also reduces the maximum transport block size, which makes data transfer less efficient.

NR allows for rapid switching between up to four BWPs using the control channel DCI. BWP switching incurs a small delay for retuning, filter settling and channel adaptation during which no data transfer is performed.

The analysis presented here assumes that 2 BWPs are configured. The first is a 20MHz BWP (power saving) that is used for control channel monitoring and the transfer of small data packets. The second is a 100MHz BWP (full power) which is used for large data packets and (while it is active) for all other purposes. When the network has large data to transmit to the UE, it uses the DCI to switch the UE to the full power BWP, where it remains while there is data to transfer. After a timeout of 20ms in which there has been no data transfer it reverts to the power saving BWP. BWP switching is assumed to add 7 slots delay when switching from a low bandwidth to a high bandwidth BWP, to allow for channel adaptation.

In the power model, a 20MHz BWP requires 40% of the power for a 100MHz BWP, subject to a floor of 50 model units.

Sleep and uplink power are not affected by BWP switching. The remaining full power BWP parameters are the same as the baseline case. For the power saving BWP, power for

PDCCH+PDSCH is 120 units (0.4*300) and power for PDCCH-only, SSB/CSI-RS and BWP switching is 50 units.

Results are presented below in Table 5-18 – Table 5-20.

FTP traffic with BWP adaptation					
Power state	Substate	Model power	Time percentage	Average power	% power
Microsleep		45	0.13%	0.06	0.15%
Light sleep		20	5.43%	1.09	2.74%
Deep sleep		1	55.04%	0.55	1.39%
PDCCH-only	Power saving	50	23.77%	11.89	29.95%
	Full power	100	8.02%	8.02	20.21%
PDCCH+PDSCH	Power saving	120	0.17%	0.20	0.51%
	Full power	300	4.05%	12.15	30.62%
SSB/CSI-RS	Power saving	50	1.28%	0.64	1.61%
	Full power	100	0.10%	0.10	0.25%
UL (0dBm)		250	0.34%	0.85	2.14%
BWP switching		50	1.67%	0.84	2.10%
Transitions	Substate	Model energy	Instances/10sec	Average power	% power
LS transitions		100	40.60	0.41	1.02%
DS transitions		450	64.32	2.89	7.29%
Saving	**23.67%**		**Total power**	**39.68**	**100.00%**

Table 5-18 FTP traffic power reduction using BWP switching

For FTP traffic, the total time spent in each power state has not changed significantly from the baseline, but a large proportion of the downlink activity now takes place in the power saving BWP. This saves 23.7% of the power. With a shorter timeout, or if the network switches the UE to the power saving BWP when it has no more data, the power savings would be higher.

IM traffic with BWP adaptation					
Power state	Substate	Model power	Time percentage	Average power	% power
Microsleep		45	0.03%	0.01	0.17%
Light sleep		20	2.88%	0.58	7.11%
Deep sleep		1	89.85%	0.90	11.09%
PDCCH-only	Power saving	50	5.15%	2.58	31.79%
	Full power	100	0.91%	0.91	11.24%
PDCCH+PDSCH	Power saving	120	0.02%	0.02	0.30%
	Full power	300	0.04%	0.12	1.48%
SSB/CSI-RS	Power saving	50	0.70%	0.35	4.32%
	Full power	100	0.01%	0.01	0.12%
UL (0dBm)		250	0.18%	0.45	5.56%
BWP switching		50	0.22%	0.11	1.36%
Transitions	Substate	Model energy	Instances/10sec	Average power	% power
LS transitions		100	21.35	0.21	2.64%
DS transitions		450	41.08	1.85	22.83%
Saving	**24.91%**		**Total power**	**8.10**	**100.00%**

Table 5-19 IM traffic reduction using BWP adaptation

In the IM case there is very little activity in the full power BWP, and the power saving is slightly higher than FTP at 24.9%. There is still significant power consumption for PDCCH-only in the full power BWP, because of the timeout. As before, a shorter timeout or an active switch back to the low power BWP would increase the power saving

VoIP traffic with BWP adaptation					
Power state	Substate	Model power	Time percentage	Average power	% power
Microsleep		45	2.99%	1.35	3.91%
Light sleep		20	43.52%	8.70	25.32%
Deep sleep		1	24.11%	0.24	0.70%
PDCCH-only	Power saving	50	22.70%	11.35	33.02%
	Full power	100	0.00%	0.00	0.00%
PDCCH+PDSCH	Power saving	120	1.25%	1.50	4.36%
	Full power	300	0.00%	0.00	0.00%
SSB/CSI-RS	Power saving	50	5.10%	2.55	7.42%
	Full power	100	0.00%	0.00	0.00%
UL (0dBm)		250	0.32%	0.80	2.33%
BWP switching		50	0.00%	0.00	0.00%
Transitions	Substate	Model energy	Instances/10sec	Average power	% power
5.322		100	322.24	3.22	9.38%
9.5589		450	103.52	4.66	13.55%
Saving	**31.63%**		**Total power**	**34.37**	**100.00%**

Table 5-20 VoIP power saving using reduced bandwidth BWP

For VoIP traffic, there are no large data packets to trigger a switch to the full power BWP, so all downlink activity takes place in the power saving BWP. This gives the best power saving of the three scenarios at 31.6%.

5.3.5.4 Antenna adaptation

Each active transceiver/antenna makes a direct contribution the UE power consumption from its own power requirement, and also an indirect contribution through the increase in processing workload. In the 3GPP model, 2 active antennas require 30% less power than 4 antennas.

The simulations for these cases use a BWP based approach with DCI based switching, as in the previous section, with a power saving 2 layer BWP for PDCCH monitoring and small data packets, and a full power 4 layer BWP for large data. Simulation results are listed in Table 5-21 - Table 5-23 VoIP traffic power saving with antenna adaptation

FTP traffic with antenna adaptation					
Power state	Substate	Model power	Time percentage	Average power	% power
Microsleep		45	0.13%	0.06	0.13%
Light sleep		20	5.40%	1.08	2.40%
Deep sleep		1	55.15%	0.55	1.23%
PDCCH-only	Power saving	70	24.61%	17.23	38.29%
	Full power	100	8.95%	8.95	19.89%
PDCCH+PDSCH	Power saving	210	0.17%	0.36	0.79%
	Full power	300	3.86%	11.58	25.74%
SSB/CSI-RS	Power saving	70	1.29%	0.90	2.01%
	Full power	100	0.10%	0.10	0.22%
UL (0dBm)		250	0.35%	0.88	1.94%
Transitions	Substate	Model energy	Instances/10sec	Average power	% power
LS transitions		100	40.42	0.40	0.90%
DS transitions		450	64.63	2.91	6.46%
Saving	**13.45%**		**Total power**	**44.99**	**100.00%**

Table 5-21 FTP traffic power saving with antenna adaptation

The power reduction for the 2 antenna case is smaller than the power reduction from 100MHz to 20MHz bandwidth, but the behaviour characteristics are similar. Downlink activity spends more time in the 2 antenna power saving configuration than it does in the 4 antenna full power configuration, achieving a 13.4% power saving. PDCCH activity at full power is still significant, but could be reduced by applying a shorter timeout, or by dynamically switching back to the 2 antenna case when there is no more data for the user.

IM traffic with antenna adaptation					
Power state	Substate	Model power	Time percentage	Average power	% power
Microsleep		45	0.03%	0.01	0.14%
Light sleep		20	2.88%	0.58	6.15%
Deep sleep		1	89.85%	0.90	9.60%
PDCCH-only	Power saving	70	5.26%	3.68	39.32%
	Full power	100	1.02%	1.02	10.89%
PDCCH+PDSCH	Power saving	210	0.02%	0.04	0.45%
	Full power	300	0.04%	0.12	1.28%
SSB/CSI-RS	Power saving	70	0.70%	0.49	5.23%
	Full power	100	0.01%	0.01	0.11%
UL (0dBm)		250	0.18%	0.45	4.81%
Transitions	Substate	Model energy	Instances/10sec	Average power	% power
LS transitions		100	21.35	0.21	2.28%
DS transitions		450	41.08	1.85	19.74%
Saving	**13.18%**		**Total power**	**9.36**	**100.00%**

Table 5-22 IM traffic power saving with antenna adaptation

The saving in the IM case is similar at 13.1%, and the FTP conclusions are also applicable to this case

VoIP traffic with antenna adaptation					
Power state	Substate	Model power	Time percentage	Average power	% power
Microsleep		45	2.94%	1.32	3.25%
Light sleep		20	43.55%	8.71	21.41%
Deep sleep		1	24.15%	0.24	0.59%
PDCCH-only	Power saving	70	22.90%	16.03	39.40%
	Full power	100	0.00%	0.00	0.00%
PDCCH+PDSCH	Power saving	210	1.20%	2.52	6.19%
	Full power	300	0.00%	0.00	0.00%
SSB/CSI-RS	Power saving	70	5.10%	3.57	8.78%
	Full power	100	0.00%	0.00	0.00%
UL (0dBm)		250	0.16%	0.40	0.98%
Transitions	Substate	Model energy	Instances/10sec	Average power	% power
LS transitions		100	322.35	3.22	7.92%
DS transitions		450	103.68	4.67	11.47%
Saving	**19.08%**		**Total power**	**40.68**	**100.00%**

Table 5-23 VoIP traffic power saving with antenna adaptation

The VoIP simulation never enters the full power substates, and the power reduction is correspondingly higher at 19%

5.3.6 Nonstandalone comparison

In this comparison the traffic model used combines the FTP and IM traffic scenarios used in previous sections, and in the NSA case routes the IM traffic through the LTE modem and the FTP traffic through the NR modem. For the SA case both sets of data traffic are routed through the NR modem

Simulation results for the LTE modem in the NSA configuration are listed in Table 5-24.
The power model parameters are based on 20MHz BWP scaling, using the assumptions of section 5.3.2

LTE modem in NSA configuration, mixed data traffic				
Power state	Model power	Time percentage	Average power	% power
Microsleep	45	0.00%	0	0.00%
Light sleep	20	2.87%	0.57	7.15%
Deep sleep	1	89.24%	0.89	11.11%
PDCCH-only	40	6.02%	2.41	29.98%
PDCCH+PDSCH	120	0.63%	0.76	9.41%
Sync/meas	40	0.63%	0.25	3.14%
PDCCH+sync/meas	68	0.31%	0.21	2.62%
PDCCH+UL	290	0.31%	0.90	11.19%
	Model energy	Instances/10sec	Average power	% power
LS transitions	100	22.34	0.22344	2.78%
DS transitions	450	40.37	1.816713	22.62%
		Total Power	**8.03**	**100%**

Table 5-24 Power consumption for the LTE modem in a nonstandalone NR configuration, mixed data traffic

The resulting power level of 8.03 is lower than the baseline standalone NR modem with IM traffic, and slightly lower than the same modem with 20MHz BWP reduction.

Simulation results for the NR modem in the NSA configuration are listed in5-25. The power model parameters are essentially the same as for the baseline case.

NR modem in NSA configuration, mixed data traffic				
Power state	Model power	Time percentage	Average power	% power
Microsleep	45	0.49%	0.22	0.43%
Light sleep	20	5.16%	1.03	2.02%
Deep sleep	1	55.41%	0.55	1.08%
PDCCH-only	100	33.95%	33.95	66.47%
PDCCH+PDSCH	300	3.51%	10.53	20.62%
SSB/CSI-RS	100	1.32%	1.32	2.58%
UL	250	0.16%	0.40	0.78%
	Model energy	Instances/10sec	Average power	% power
LS transitions	100	38.93	0.39	0.76%
DS transitions	450	59.60	2.68	5.25%
		Total Power	**51.08**	**100%**

Table 5-25 Power consumption for the NR modem in a nonstandalone NR configuration, mixed data traffic

In this case the power level of 51.08 units is lower than for the baseline standalone NR modem with FTP traffic (51.99 units). The power distribution is very similar to the baseline case, but there is a slight reduction in PDCCH+PDSCH traffic for the NSA case, which may be due to signalling differences

Results for the standalone NR modem are listed in Table 5-26.

NR modem in standalone configuration, mixed data traffic				
Power state	Model power	Time percentage	Average power	% power
Microsleep	45	0.51%	0.23	0.43%
Light sleep	20	5.03%	1.006	1.88%
Deep sleep	1	53.42%	0.5342	1.00%
PDCCH-only	100	35.91%	35.91	67.15%
PDCCH+PDSCH	300	3.67%	11.01	20.59%
SSB/CSI-RS	100	1.30%	1.3	2.43%
UL	300	0.16%	0.48	0.90%
	Model energy	Instances/10sec	Average power	% power
LS transitions	100	37.99	0.379857	0.71%
DS transitions	450	58.31	2.6239275	4.91%
		Total Power	**53.47**	**100%**

Table 5-26 Power consumption for a standalone NR modem, mixed data traffic

Compared with the baseline FTP case (51.99 units), the additional IM traffic means that power for the mixed traffic case increases to 53.47 units.

The combined power for the nonstandalone mixed traffic case is (8.03+51.08) = 59.11 model units, compared with the standalone mixed traffic power of 53.47 model units. For this traffic pattern the UE in the NSA configuration requires 10.55% more power from its battery than in an SA configuration.

6 Requirement of 5G terminal power consumption

6.1 NSA

In a non-standalone deployment the UE is connected to a master node which is an eNB (EN-DC) or an ng-eNB (NGEN-DC), and communicates with it using E-UTRA protocols [2]. When the UE is in the connected state, NR connectivity is provided by a secondary node which is a gNB. Other forms of dual connectivity exist, but are not considered here. The UE RRC state is derived from the master node RRC state.
Dual connectivity inevitably increases power consumption compared to single connectivity, so for the lowest UE power consumption in an NSA deployment the secondary node connection should be activated only during periods of high data traffic.

6.1.1 RRC idle state and RRC inactive state

6.1.1.1 RRC idle state

The UE is in the idle state when there is no RRC connection established. In RRC_IDLE:

A UE specific DRX cycle may be configured by upper layers
UE monitors a paging channel for core network paging using 5G-S-TMSI or IMSI
UE performs neighbour cell measurements and cell reselection
UE acquires system information (SI)

The UE uses neighbour cell measurements to manage its mobility. If cell reselection to a different tracking area is required it notifies the network using a RACH procedure. SI acquisition and cell reselection are assumed to occur sufficiently infrequently that their impact on UE average power consumption can be neglected.

The DRX cycle defines how often the UE needs to wake up to check for paging (320, 640, 1280 or 2560ms). During the wakeup it also measures the signal strength of the serving cell, and if this is below the cell reselection criteria it initiates measurement of the neighbour cells. Since there is an energy cost to waking and returning to sleep, these activities are usually performed in the same wakeup as the paging occasion.

The wakeup duration is generally small as a proportion of the DRX cycle, and average power in this state is dominated by the sleep state power and the energy cost of waking and returning to sleep.
Measurements on neighbour cells can be scheduled in the same wakeup as the paging occasion to reduce power consumption, but this may increase the wakeup duration if the symbols needed for the measurement only occur infrequently, or if measurement of inter-frequency cells requires RF retuning. For an NSA deployment, it is assumed that idle mode measurements will not include the NR carrier because reselection to the NR cell is not possible. This means that the UE has no knowledge of the secondary gNB before entering the connected state, and would need to complete cell search and beam alignment procedures before it could access the secondary node.

All activities are performed in the master E-UTRA cell, with the exception of neighbour cell measurements on other E-UTRA cells, so power consumption in this state is expected to be similar to the equivalent power for an E-UTRA UE.

6.1.1.2 RRC inactive state

The RRC_INACTIVE state is not present in EN-DC. It only exists in NGEN-DC when the master node is an ng-eNB.
The UE is in the inactive state when there is an RRC connection which has been suspended. In inactive mode the UE performs all of the functions that are part of the idle state, but additionally in RRC_INACTIVE:

- A UE specific DRX cycle may be configured by the RRC layer
- UE Stores the access stratum context
- A RAN-based notification area is configured by the RRC layer
- UE performs RAN-based notification area updates periodically
- UE performs additional updates when moving outside the configured RNA
- UE monitors a paging channel for RAN based paging using I-RNTI

A UE in the inactive state can move between nodes within its notification area without needing to inform the network. The network is only notified when the UE moves outside its notification area.
The RAN-based activities are not expected to add significantly to the wakeup energy compared with idle mode, but in the inactive state a shorter DRX cycle may be configured to give a faster response, and additional measurements would be performed (including the NR cell(s)). A shorter DRX cycle would result in higher RRC_INACTIVE power consumption than for RRC_IDLE.
RAN-based paging initiates an RRC Connection Resume procedure which returns the UE to the RRC_CONNECTED state. CN paging would return the UE to the idle state and inform the NAS that paging had occurred.
The network has choices to make in how the RRC_INACTIVE state is managed. If a similar DRX cycle is configured for both RRC_INACTIVE and RRC_IDLE, the difference in power consumption between the two will be small, and keeping the UE in RRC_INACTIVE is beneficial in giving faster transitions to the RRC_CONNECTED state.

If alternatively a faster DRX cycle is configured for RRC_INACTIVE to improve response times there will be a significant increase in power consumption since the UE spends more time monitoring the control channel. In this case the network should configure a timeout, such that if the time spent continuously in RRC_INACTIVE without a transition to RRC_CONNECTED exceeds the timeout, the UE will revert to the RRC_IDLE state
If these principles are followed, time spent in the RRC_INACTIVE state is not expected to make a large contribution to UE average power consumption, but it offers the advantage of a quicker transition to the connected state than is available in idle mode, and provides lower power consumption than connected DRX.

6.1.2 RRC connected state

In the connected state the UE monitors the control channel for indications that user data or signaling is present. In an NSA deployment this monitoring can take place on both RATs. Additional measurements are performed to assess channel quality, and the results are reported to the network periodically. In the connected state mobility is managed by the network based on the measurements reported by the UE.

6.1.2.1 No data transmission without DRX or with DRX

If there is no data traffic and DRX is not enabled, the key to lower UE power consumption is to reduce the power required for PDCCH reception and decoding. Monitoring only on the master node would be the easiest way to achieve this, but quick activation and deactivation of the secondary node would be needed to make good use of the NR data capability, and the additional delays would impact on latency. Dual-RAT PDCCH monitoring would increase monitoring power consumption significantly, with two independent monitoring paths active instead of one, so there may be power saving benefits from cross-carrier scheduling.

The power increase from dual-RAT monitoring can be mitigated by using a low-duty cycle DRX configuration, but in a dual-RAT scenario DRX operation has the potential to be more complicated and less efficient, as there is no simple mechanism in the specifications for synchronising the DRX cycles between the master and slave nodes. For this reason it is highly desirable that operators of NSA deployments ensure DRX cycle boundaries are aligned between the E-UTRA and NR cells to reduce the number of separate wakeup occasions.

An important power saving provision currently under discussion in 3GPP is the concept of a dormant state for the SCell. In the dormant state measurements would take place on the SCell for CSI, RRM and beam management, but there would be no PDCCH monitoring on the cell until it was activated. This would mean that the SCell could be dormant during no-data periods, but activated quickly when high data bandwidth is needed.

6.1.2.2 Uplink data transmission in NR side only (high data rate and low data rate)

When there is uplink data to transmit, the node which has the lowest path loss should ideally be selected, since this will require the lowest transmitted power per subcarrier, and give the lowest power consumption at the UE. Uplink allocations can be as little as 1 symbol, but at high data rates it is likely that multiple symbols will be required for each transmission. In most cases transmission over NR would use less energy in total, as the duration of the transmission would be shorter than in E-UTRA.

If the NR SCell is dormant, there will be signaling costs for activation and deactivation to allow transmission to take place, so before activating the secondary node there should be sufficient data in the uplink, downlink or both to justify these costs.

At low data rates, if path losses are similar, there is little reason to prefer one over the other for reasons of power consumption, and if the SCell is not already active, transmission over the E-UTRA master node would be preferable.

6.1.2.3 Uplink data transmission in NR side and LTE side (high data rate and low data rate)

In NSA configurations there is generally a disparity in bandwidth between the E-UTRA node and the NR node such that the NR node has significantly higher capacity than the E-UTRA node in the uplink. There is also an upper limit on the total power that a UE can transmit, so increasing the number of uplink carriers limits the power that each can transmit.

This means the uplink coverage is generally better for one carrier transmitted at full power than for two carriers transmitted at half power. If the data can be transmitted over a single node it is also more power efficient for the UE to process one grant and one acknowledgement rather than two, and it means that the wakeup period is only extended in one DRX cycle rather than two.

At high data rates it is obvious that the single node should be the NR node, but at low data rates the signaling cost of activation and deactivation needs to be taken into account.

There will clearly be occasions where control is transmitted to the master node and data to the secondary node, and also high data rate transmission scenarios where the network dedicates the resources of both nodes to a single user to provide faster throughput, but the asymmetry between the two carriers means that energy consumption per bit at the UE is likely to be higher in such cases. Carrier aggregation works best if the carriers are of similar bandwidth.

6.1.2.4 Downlink data transmission in NR side only (high data rate and low data rate)

UE downlink efficiency (energy consumed per bit received) improves as the bandwidth and MIMO configuration increase, but so does the total power consumption. Unlike the uplink, data is received across the entire Rx bandwidth rather than just the resource block allocation, so data transfer efficiency falls when the RB allocation is a small fraction of the available bandwidth due to the energy spent processing the unallocated RBs. It is therefore desirable to adapt the Rx bandwidth to the volume of data traffic.

In an NSA deployment, there is already a low bandwidth E-UTRA channel and a high bandwidth NR channel, so UE power consumption benefits if the network routes small transport blocks via the master node and larger transport blocks via the secondary node.

BWP for the NR SCell can be configured to adapt the NR Rx bandwidth to the data size, but if BWP selection is done via the DCI to improve speed it is also necessary to configure cross-slot scheduling. For NSA, if the SCell is activated only during periods of high data traffic the bandwidth occupancy should be high, and the savings from this type of adaptation may therefore be small as a proportion of total UE energy usage. The effect would be greater in a standalone deployment.

If high data rates (in either the downlink or the uplink, or both) are sustained for prolonged periods, the UE may experience overheating. If this occurs, the UE should request a temporary capability reduction from the network. This will allow it to disable some resources, or operate them at lower power to maintain a safe temperature. Details of signalling for the request are still under discussion.

6.1.2.5 Downlink data transmission in NR side and LTE side (high data rate and low data rate)

Because there is a power penalty for dual connectivity, UE downlink power consumption in NSA deployments is also minimised by activating the NR connection only during periods of high bandwidth traffic and using the E-UTRA connection when data rates are lower.

Some control signaling has to take place via the master node, but simultaneous routing of data traffic to both nodes should be avoided where possible, as this will extend the active period of DRX in both of the DRX cycles, whether or not they are synchronised. This will increase the total wakeup time, and therefore the power consumption.

6.1.2.6 Voice

To deliver good quality bidirectional voice traffic in real time, as required for voice telephony, the UE needs to send and receive speech packets each representing 20ms of audio with an end-to-end delay of less than 100ms. To achieve this in a power efficient way, the UE needs a DRX cycle of 20 or 40ms, with short OnDuration and inactivity timers. The data rate is low (kilobits per second), and can easily be carried by the master or the slave node.

Semi-persistent scheduling of resources can help to reduce the need for PDCCH monitoring – for voice traffic the uplink and downlink speech packets and their respective ACK/NACK transmissions should be scheduled close together in time to minimise the wakeup duration

If voice traffic is the only UE activity, it is difficult in an NSA deployment to justify activation of the secondary node, as this will increase power consumption on the UE. It may be necessary if there is congestion at the E-UTRA node, but if higher data rate traffic is routed preferentially over NR this should not be a frequent occurrence.

6.2 SA

In a standalone deployment the UE is connected to a master node which is a gnB, and communicates with it using NR protocols [2]. When the UE is in the connected state, there may additionally be one or more secondary nodes which can be gNB or en-gNB. The UE RRC state is derived from the master node RRC state.

An SA deployment provides a high data bandwidth connection without carrier aggregation or dual connectivity, although additional SCells can be activated to increase peak throughput.
For reasons already discussed, this should only be done when data requirements are particularly intense.

A UE with low data activity connected to an SA network is operating in a high bandwidth environment, and potentially has to discard a high proportion of the downlink data that it processes. This could be bad for power consumption, but NR has introduced a number of new features which can be configured by the network to improve UE power efficiency when data rates are lower.

These features can also be applied in non-standalone networks, but are discussed in more detail here because the UE in an NSA network is assumed to be connected to the NR secondary node only when high bandwidth is needed.

6.2.1 RRC idle state and RRC inactive state

6.2.1.1 RRC idle state

In an SA deployment the idle state functions of the UE are very similar to those listed in 6.1.1.1, but paging is only by 5G-S-TMSI, and the UE may send SI requests to the network. Neither of these differences has a significant effect on power consumption.

However, in an SA configuration measurements for cell reselection can include potential secondary nodes, and if this is the case then transitions to the RRC_CONNECTED state can be much faster than for the NSA case.

It should be noted that a paging occasion in an SA master node may have a longer duration than for an NSA master, because in an NR cell the paging duration corresponds to the beam sweeping period [4]. The selection of beams for paging and measurement is up to the UE implementation, but it is reasonable to suppose that if the configuration of measurement resources means a longer wakeup for the UE there may be some increase in the RRC_IDLE average power.

Initial modelling suggests that the average power in RRC_IDLE is likely to be dominated by the sleep power and the energy cost of the sleep-wakeup transitions, so for similar measurement configurations any net increase compared to NSA is likely to be small.

6.2.1.2 RRC inactive state

The RRC_INACTIVE state is a new feature in NR, and is not present in E_UTRA. The discussion in section 6.1.1.2 is therefore equally applicable to SA deployments.

6.2.2 RRC connected state

6.2.2.1 No data transmission without DRX or with DRX

In an SA deployment there are opportunities for PDCCH-only power reduction that are only available when the master node is NR (similar techniques can be employed in the secondary node in NSA, but are less effective if the secondary node is dormant most of the time).

One important technique for reducing power in no-data scenarios (new in NR) is cross-slot scheduling. In E-UTRA, the control channel indicates the presence or absence of data in the same slot in which it is received, which means that the receiver must remain active while the control channel is being decoded in case data is present. In NR, the control channel in slot N indicates the presence of data in slot N+K0, where K0 is a configuration parameter.

K0 is 0 by default, but if the network configures a larger value the UE can disable its receiver as soon as PDCCH reception is complete, thereby reducing power consumption. PDSCH reception then only takes place in slots where there is data present for the UE to decode.

The NR power model recently agreed in 3GPP TR38.840 [13] suggests that PDCCH-only power will be approximately 1/3 of the power for peak downlink throughput with K0=0, and that cross-slot scheduling can reduce this by a further 30%.

Reducing the frequency of PDCCH monitoring (while observing latency requirements) would also reduce UE power consumption. This is currently under discussion for Rel-16, and would allow multiple slots of PDSCH to be scheduled from a single PDCCH slot, increasing resource utilisation.

Cross-slot scheduling and PDCCH monitoring reduction both increase latency in data traffic, but trading latency for power reduction will be acceptable for most users. The default configuration should favour low power consumption, with applications that require low latency from the network requesting it explicitly.

Bandwidth part adaptation is another power saving feature that is new to NR. By reducing the bandwidth that is captured and processed, the baseband throughput can be significantly reduced, giving further power reductions of up to 60% [3] Reducing the number of receive paths provides additional opportunities for power saving.

The reduction in throughput with BWP adaptation and Rx path reduction applies to the data channel as well as the control channel – switching to a higher bandwidth may be needed when there is more data present, and rapid switching of BWP via the DCI is desirable to avoid unnecessary increases in latency.

6.2.2.2 Uplink data transmission (high data rate and low data rate)

In the uplink, Tx power control determines the transmitted power in each subcarrier, and the RB allocation is matched to the size of the transport block. The network will normally configure the highest order modulation scheme that the channel can support, so there are few opportunities for reducing UE power consumption in uplink transmissions.

Reducing the number of transmitting antennas from 2 to 1 can save UE power in some scenarios, but this decision depends on channel conditions, and should take into account the throughput benefits of diversity gain or additional MIMO layers.

In range limited scenarios there is an upper limit on the size of RB allocation that can be supported, due to the maximum power that the UE can transmit. Reducing the uplink duty cycle can allow the UE to increase its transmit power provided that the average power remains within the limit, and if necessary, repetition can be used to increase the range of coverage.

6.2.2.3 Downlink data transmission (high data rate and low data rate)

In the downlink, for a given subcarrier spacing, processing requirements increase with bandwidth and the number of receive paths. A larger subcarrier spacing reduces the processing requirement, but shortens the TTI so that processing must be completed more quickly.

Higher bandwidth can deliver data faster with lower energy per bit to more users, but from a UE perspective decoding of resource blocks that are not part of its allocation constitute wasted energy. Energy consumption is reduced if data transmission occupies many resource blocks and few TTIs compared to occupying few resource blocks in many TTIs in a narrow bandwidth. This principle also applies when transport block sizes are small, but if the processed bandwidth is substantially larger than the UE throughput, energy is wasted.

These fundamental principles become more and more important as the bandwidth of the channel increases. The ultimate limit is set by thermal noise, and as bandwidth grows so does the noise within the channel. Path delay spread leads to an increase in subcarrier spacing, which increases the noise per subcarrier. In order to design a high bandwidth network that is power efficient, it is essential that no node in the network has to process significant quantities of irrelevant data.

This implies that capture of the data channel should be avoided unless the control channel indicates that relevant data is physically present. This in turn implies that there should be a delay between reception of the control channel and reception of the data channel to which it refers. This delay should be sufficient for the UE to configure the bandwidth of data reception to match the transport block size.

The success of NR (and its successors in mobile communication technology) will depend on how well these principles are embodied in the network structure. Cross–slot scheduling, bandwidth adaptation, transceiver path reduction, and DRX cycles that are adapted to traffic patterns and latency requirements will be key elements in ensuring that user devices in a mobile communications network operate in the most power efficient manner.

UE overheating remains a possibility if the highest data rates are maintained for long periods, and signaling will be necessary to reduce the UE throughput if this occurs

6.2.2.4 Voice

Voice telephony requires only a fraction of the bandwidth offered by NR, so in an SA deployment a narrow bandwidth part is a basic requirement for low UE power consumption. Reducing the number of antennas for uplink and downlink traffic would allow further savings. Data traffic in both directions is predictable, with speech packets generated every 20ms, so semipersistent scheduling reduces the need for control signalling. A DRX cycle of 20 or 40 ms matches the data traffic pattern and meets the end-to-end latency requirement.

For best power consumption the on duration in the DRX cycle should be short – combining the uplink speech packet with the downlink acknowledgement (or vice-versa) can reduce the number of data transactions per cycle to 3, and mini-slot would allow these to be scheduled in close proximity to each other, but provision must also be made for background measurements and possible retransmissions. The distribution of measurement resources may limit how short the wakeup duration can be - it is not possible to schedule every user adjacent to an SSB transmission.

6.3 Estimating power consumption in NR UEs

The NR power model agreed in 3GPP [3] describes how UE power consumption scales in relative terms for different configurations, but because it represents a consensus between several UE suppliers there is no absolute calibration of the power unit in the model. The model does not include E-UTRA operation, so it is only applicable to the standalone configuration.

CMCC have produced a requirements specification for the sub-6GHz NR UE , which includes targets for UE power consumption in different configurations. UEs that are approved for use on its network are expected to meet target performance, so these targets can be viewed as providing a pessimistic estimate of UE power consumption. Table 6-1 below compares some key model parameters to the power consumption requirement of the nearest related configuration in the CMCC specification

Table 6-1 Comparison between the 3GPP model and CMCC requirement

3GPP model [3]		CMCC requirement	
Slot Configuration	Model power	Power (mW)	Configuration (TDD 7D1S2U)
PDCCH+PDSCH	300	2400	SA 100MHz 1CC, 1Gbps DL, 0dBm Tx
UL transmission 0dBm	250	1880	SA 100MHz 1CC, 100Mbps UL, 0dBm Tx
UL transmission 23dBm	750	3600	SA 100MHz 1CC, 100Mbps UL,23dBm Tx

It is clear from the table that the power scaling is different in the comparison cases, but the comparison is not direct, because data transfer in either the uplink or the downlink requires a mixture of transmissions in both directions. Noting that the CMCC configurations allocate only 20% of slots to uplink transmission, reasonable alignment between the two sets of data can be obtained if we assume a conversion factor of 1 model unit approximately equal to 8mW.

Based on the CMCC specification, we can expect a UE in an NSA deployment to require 15-25% more power when active and transferring data in the NR cell only. If both RATs are used for data transfer the power penalty would be higher, but so would the data rate.

At maximum throughput and maximum transmit power, the allowable power consumption is at a level where UE overheating becomes a possibility, and for the HPUE case even higher instantaneous power levels are possible, although these cannot be sustained on a continuous basis. For most use cases these high throughput conditions only occur in short bursts, but if substantial temperature rises occur, UE assistance signalling may be needed to request a reduced power configuration.

However, these are extreme cases – more typically, UEs will spend a high proportion of their time in the idle or inactive state, where power consumption is much lower, and even in the connected state a UE will spend much of its time monitoring the control channel waiting for data to arrive. In the 3GPP model this PDCCH-only state (operating continuously at maximum bandwidth) consumes only 1/3 of the power required at peak throughput.

PDCCH-only power (estimated at 800mW using the previously derived conversion factor) can be reduced even further by network configuration. Based on 3GPP model assumptions, reducing the control channel monitoring bandwidth from 100MHz to 20MHz can deliver a 60% reduction in monitoring power, with a further 30% saving (combined saving of 72%) if cross-slot scheduling is enabled. In delay-tolerant applications DRX can provide additional power reductions by reducing the duty cycle for monitoring activities, bringing average power consumption down to very low levels.

This brief overview suggests that instantaneous power consumption in an NR UE can vary over a wide range in normal use, but if the network configuration allows the UE to perform demand-based switching between efficient high bandwidth data transfer, reduced bandwidth low power control channel monitoring and periods of inactivity then NR can deliver a user experience that combines rapid responsivity and throughput with good battery life.

7 Service

7.1 Service type and Parameter configuration

- Note：Model analysis, parameter configuration and power consumption effect

To evaluate the metric of different power saving schemes, diverse traffic with different characteristic need to be considered. According to the RAN1#94bis agreement [3], FTP model 3 should be included in the evaluation for at least FTP application. Other bursty traffic arrival models can also be considered. In addition, other applications including web-browsing, video streaming, instant messaging, VoIP, gaming and background app sync can be considered for traffic modelling for power saving proposal evaluation.

7.1.1 FTP model 3

FTP models already defined in in 3GPP TR 36.814 can be at least a starting point for traffic modelling. It is characterised as large packets size, e.g., Mbytes and less frequent packet arrival, e.g., hundreds of ms or several seconds. FTP model 3 (use 0.1 Mbytes packet size, mean inter-arrival time 200msec) is used for the purpose of basic calibration of traffic modeling.

7.1.2 Game

Gaming is characterised as smaller packets size, e.g., hundreds of bytes, frequent and random packet arrival, e.g., tens of ms. The relevant parameters of the gaming traffic are as follows.

Table 7-1 Parameters for gaming

Parameter	Statistical Characterization
Packet arrival	For packet arrival of <60ms, fixed probability of 0.4%, i.e., packet arrival for any value from 0 to 59ms has fixed 0.4% probability; For packet arrival of >= 60ms, Largest Extreme Value Distribution (also known as Fisher-Tippett distribution), PDF: $f_x = \frac{1}{b} e^{-\frac{x-a}{b}} e^{-e^{-\frac{x-a}{b}}}$, x is the packet arrival (in ms) and f_x is the probability of x. a= 66 and b= 3. Without loss of generality, a= [30-80] and b= [2-5] can be considered. Values for Fisher-Tippett distribution can be generated by the following procedure: x = a – b ln (– ln y), where y is drawn from a uniform distribution in the range [0,1]
Packet size	Largest Extreme Value Distribution (also known as Fisher-Tippett distribution): PDF: $f_x = \frac{1}{b} e^{-\frac{x-a}{b}} e^{-e^{-\frac{x-a}{b}}}$, x is the packet size (in Bytes) and f_x is the probability of x. a= 220 and b= 25. Without loss of generality, a= [200-300] and b=[20-30] Values for Fisher-Tippett distribution can be generated by the following procedure: x = a – b ln (– ln y), where y is drawn from a uniform distribution in the range [0,1]

7.1.3 Video

The typical video streaming traffic (i.e. YouTube) modelling is described with two phases, an initial burst phase followed by a throttling phase (Figure 7-1). In the initial phase, the video streaming progressive download commences by transferring an initial burst of data with several seconds. And then in the throttling phase, the video server throttles down the traffic generation rate during which the traffic pattern alternates between the reception of data chunks and short periods without packets. It is illustrated in the following figure.

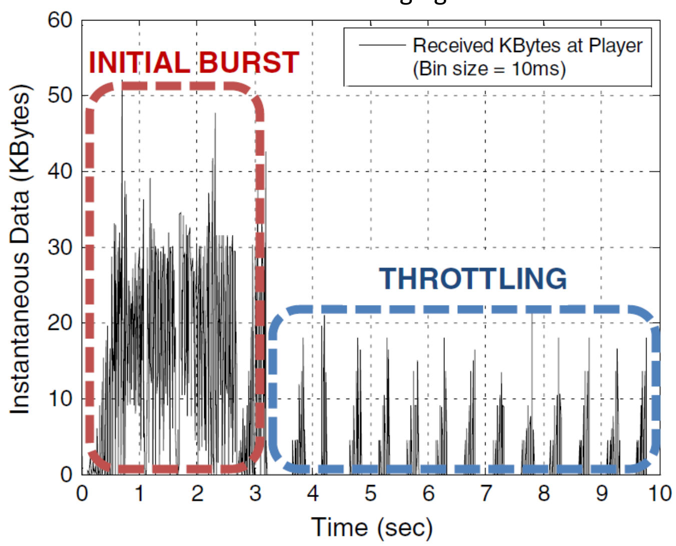


Figure 7-1 Initial burst phase and throttling phase in video streaming traffic modelling

Since the initial phase of video streaming more looks like the full buffer traffic, it is sensible to only model the throttling phase for power saving traffic.

7.1.4 Voice

Data generated by voice calls are highly deterministic, as the corresponding codecs generate data at known intervals and of a known size. The relevant parameters of the VoIP traffic are as follows from R1-070674.

Table 7-2 Parameters for VoIP

Parameter	**Characterization**
Codec	RTP AMR 12.2, Source rate 12.2 kbps
Encoder frame length	20 ms
Voice activity factor (VAF)	50% (c=0.01, d=0.99)
SID payload	Modelled 15 bytes (5Bytes + header) SID packet every 160ms during silence
Total voice payload on air interface	40bytes (AMR 12.2)

7.1.5 Other Applications

Additionally, some other representative applications can be selected. The following applications have been mentioned such as web-browsing, instant messaging, and background app sync.

For web-browsing application, it is usually modelled to request several web pages with different inter-arrival time for user reading the web page. A web page (web session) consists of a main object following some other inline objects. After the downloading process of one web session, a session inter-arrival period takes place. Web-browsing can use HTTP to model.

Table 7-3 Web Browsing traffic modelling parameters

Model Parameters	Description	Distribution
Session inter-arrival time	Gaps between sessions (User reading time)	Weibull
Number of Objects per session	Number of object per session	Gamma
Object inter-arrival time	Delay between the arrival of two objects	Gamma
Object size	Size of each object	Weibull

For instant messaging, the traffic tends to be sporadic. For simplicity, the inter-arrival time also can be modelled as fixed value. FTP Model 3 with some modifications can be used to model instant messaging applications.

For background app sync application, for power consumption evaluation purpose, it can be assumed that idle mode operations (inclusive of page detection, RRM, deep sleep and transition overhead) contributes to X% of the use case power. The remaining portion is contributed by intermittent RRC connections due to background activities (FFS: value of X). In summary, the traffic characteristics are captured in the following table.

Table 7-4 Proposed traffic models and related characteristics

Application	Characteristics
FTP model 3	Large packets size, e.g., Mbytes Less frequent packet arrival (exponential distribution), e.g., hundreds of ms or several seconds
Gaming	Smaller packets size, e.g., hundreds of bytes Frequent and random packet arrival, e.g., tens of ms
Video streaming	Initial burst size, throttled data rate and chunk size.
VoIP	Small packets size, e.g., tens of bytes Shorter and fixed packet arrival, e.g., tens of ms Asymmetric in UL and DL traffic
Web-browsing	Session inter-arrival time, Number of objects per session, Object inter-arrival time, Object size
Instant messaging	The size of the instant message is determined by the Pareto distribution. The inter-arrival time between two messages is modelled as the Lognormal distribution
Background app sync	RRC connections every few minutes. May be modelled as a fixed power cost relative to I-DRX

According to the RAN1#94bis agreement, percentage power consumption reduction from the baseline scheme will be used to express the power saving gain. And Latency of packet or scheduling delay, user throughput should also be reported as the result of the evaluation, in addition to power saving gain. On UE side, it means support of good power saving capabilities while at the same also provide for good user experience in terms of delay. On network side it includes support of flexible scheduling possibilities of UE's in Active mode while being able to provide good power saving possibilities to UE's in Active mode.

DRX has been used for UE power saving since Rel-99 UMTS and continued in LTE and NR by allowing UE to get into deep/light sleep state during the DRX OFF period. UE would autonomous wake up before DRX ON cycle in preparation for the signal processing. However, most of time, UE wakes up at the DRX ON period and gets no grant from PDCCH and no data from PDSCH especially for the sporadic services like instant messaging traffic. PDCCH-only state is the highest contributor.

Possible solution is the introduction of WUS (Wake up signal) which allows UE to skip the upcoming on duration period and go back to sleep. Go-to-sleep signaling can be used to stop drx-onDurationTimer or drx-InactivityTimer for faster skipping of monitoring.

Another solution to improve this condition is to enhance the current DRX, i.e., more dynamic and flexible C-DRX configuration can be supported. gNB can make decisions and change the C-DRX configuration more dynamically by some Layer1 signalling. In addition, UE can give a recommendation of the DRX configuration according to its traffic characteristics and power saving requirements.

7.2 User model

Based on different user model to analysis the device last time

7.3 5G industry application power analysis

7.3.1 7.3.1 Laptop

- Need define power consumption target per running status :
  - Power off
  - Standby
  - Idle
  - Connected standby (w/ WiFi or WWAN)
  - Local Video playback (720p, 1080p, 2k, 4k)
  - Video streaming (both WiFi and WWAN).
  - 3DMark – full performance Power.
- Need define power target per system configure :
  - Storage type and size (SSD or HDD, 128GB to 1TB, etc.
  - Memory type and size (DDR3, DDR4, LPDDR3/4/5, 2GB – 32GB, etc.).
  - Discrete display card or integrated display card
  - CPU model and power rate
  - Other factors ……
- Need have enough range to accommodate different type of laptop with different configurations.
- Typical power consumption requirement (Subject to PC configurations – Display, CPU, memory, etc.) :
  - Local video playback (1080p): 8 hours.

- Web browsing: 5 hours.
- Video streaming (WIFI / 5G-WWAN): 5 hours / TBD
- MM(Mobile Mark) battery life : > 5 hours

8 Power consumption test

8.1 Test instrumentation

The specification relates to test instruments including,

1. Common Test System Architecture

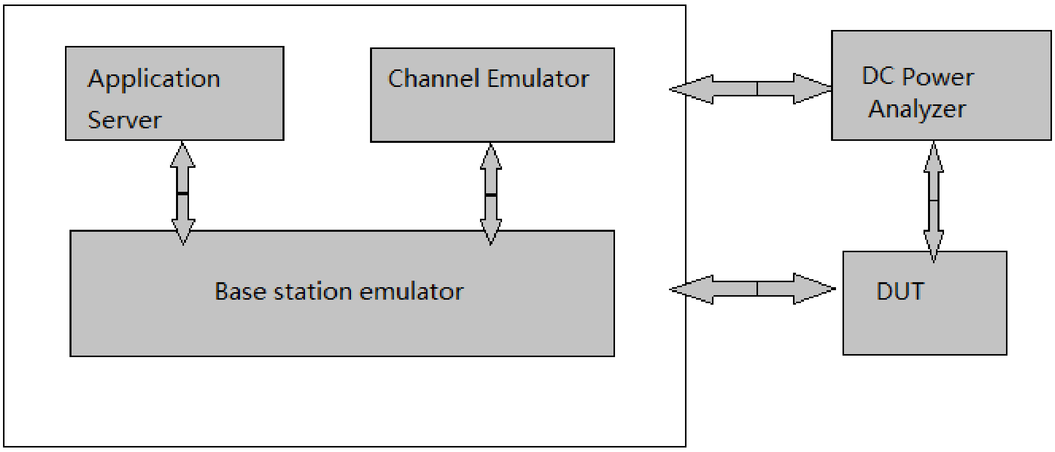


Figure 8-1 Common Test System Architecture

2. System Simulator including Application Server and Base Station Emulator. Following function should be considered.

    - Frequency bands and System bandwidth:

Table 8-1 5G NR Band list

RAT	Band	Max BW for 15kHz	Max BW for 30kHz	Max BW for 60kHz
NR	Band n77(3.3GHz ~ 4.2 GHz)	50MHz	100MHz	100MHz
	Band n78(3.3GHz ~ 3.8 GHz)	50MHz	100MHz	100MHz
	Band n79(4.4GHz ~ 5GHz)	50MHz	100MHz	100MHz
	Band n1	50MHz	20MHz	20MHz
	Band n3	50MHz	30MHz	30MHz
	Band n8	50MHz	20MHz	NA
	Band n41	50MHz	100MHz	100MHz

    - MIMO configurations:
        - UL: 2 layers required, 4 layers recommended
        - DL: 4 layers required, 8 layers recommended
    - NSA(Option 3X) and SA(Option 2).
    - Full stack including PHY/L2/RRC/NAS compatible with 3GPP Release 15.
    - UDP/IP data transmission service.
    - Voice over 5G NR(TBD).

3. DC Power Analyzer includes following functions:
    - High accuracy (detailed requirement is TBD).
    - Data logger function.
    - Multiple outputs.

4. Channel Emulator(detailed requirement is TBD)

8.2 Test method

1. **Test Environment:** Refer to the Test System Architecture in 8-1, the test environment would be simulated by based station emulator and Channel Emulator. Following parameter and scenarios should be considered:

- **Test Frequency:** should refer to 3GPP Test Specs, such as TS38.521-1.
- **DL power level:** it emulates the network coverage situation. A user on the edge of a cell will inevitably experience higher power consumption than one who is near a gNB.
- **UL power level:** The uplink power will vary to simulate the distance between the UE and the gNB. And uplink power is dominated by the UE transceiver contribution.
- **Bandwidth part adaptation:** It can be used to reduce UE power consumption when the full data bandwidth of NR is not required.
- **Setting for Cross-slot scheduling and DRX:** Cross-slot scheduling can be used to eliminate unnecessary capture of the data channel in slots where no data is present for the UE. While DRX provides a means of trading latency for further power reductions. Used in combination, these two features can reduce the energy cost of control channel monitoring In many common low data rate use cases control channel monitoring can consume more power than data transfer, so this can lead to significant power reductions.
- **Multi-carriers and MIMO setting:** The UE transceiver power is a function of the number of active carriers and the number of MIMO layers supported on each.
- **User data throughput:** The simple power model presented suggests that NR UE power consumption at peak throughput may be higher than a LTE UE offering lower throughput, but the energy per bit at maximum performance will be significantly better.
- **User patterns:** Actually, most of real use case for NR mobile phone would be lower throughputs. So, the key to competitive NR UE performance will be how well the power consumption scales with data rate at lower throughputs. This will require cooperation between networks and UE to match UE resource allocation to data traffic patterns in ways that allow the UE to reduce the data bandwidth that it has to process, and the time for which it is active, so that it can optimize its power consumption.
- **Channel condition:** If the increase in throughput is proportionately greater than

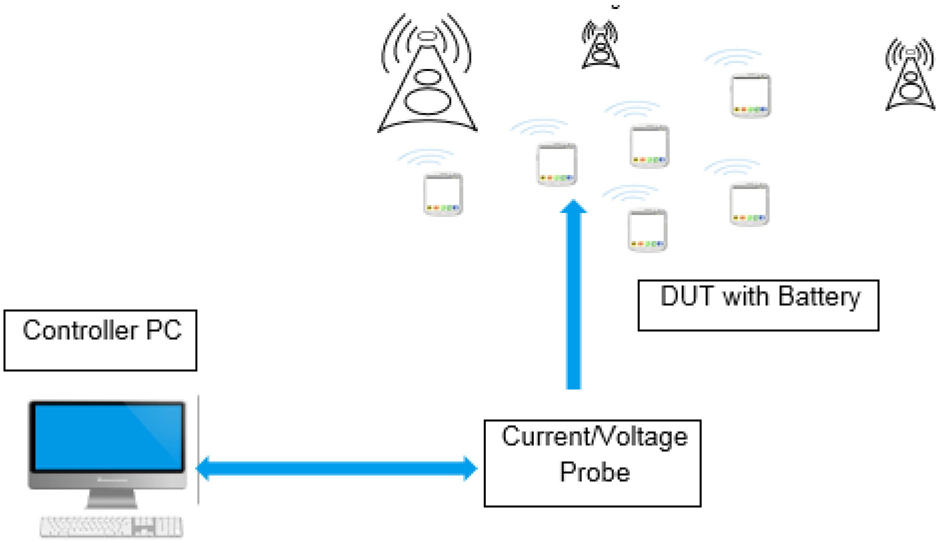
the increase in battery power then there will be a net energy saving, but this tradeoff will depend on channel conditions.

- **NSA/SA network configuration:** In principle a standalone UE should consume less power than a non-standalone UE, requiring fewer receive and transmit resources for equivalent bandwidth. However, in some case, Synchronizing DRX wakeups between LTE and NR will make NSA operation more power efficient.
- **Voice service on NR:** As one of low data rate services, the key to low power consumption for NR voice will be to keep the data that the UE has to process for a voice call to a minimum.

2. **Test Procedure:** Refer to the Test System Architecture in 8-1, the general test procedure would be followings:
   - Connect RF ports of DUT and based station emulator.
   - Connect Power ports of DUT and power analyzer.
   - Initial based station emulator and channel emulator to pre-defined parameter and settings.
   - DUT power on and perform pre-defined scenarios.
   - Record test results by power analyzer.

3. **Real network real battery DUT power consumption test**

   To maximally simulate the real working situation of DUT in real network, and evaluate the DUT efficient work time, it is important to test power consumption in real network with standalone battery.



DUT working with standalone battery without DC power supply.

Current/Voltage probe, for battery-powered devices, probing can be done at the battery connectors.

Controller PC could control and read/record the current/voltage value.

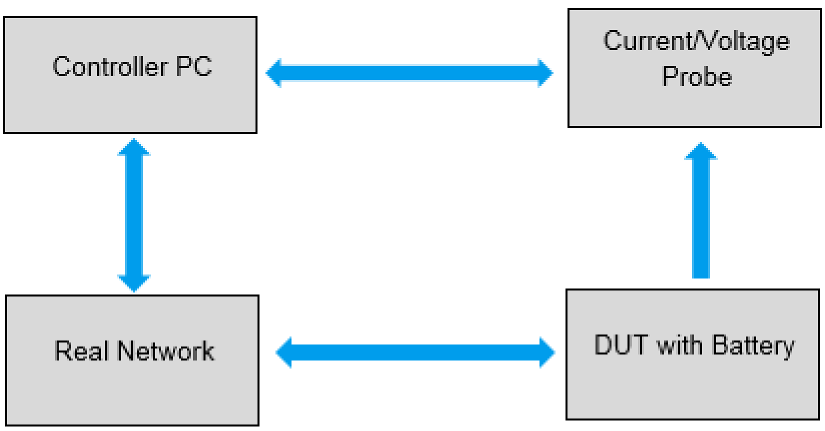
The beginning condition of battery should be defined as precondition like 100%, 50%, 20%.

The test location should include near point, middle point and far point from gNB

The test situation should include data throughput, voice and so on.

4. Real battery lab power consumption test

To make it more easily and test the real battery in lab, the following test system could be used to test real battery power consumption in lab.



- Network simulator parameter configuration refer to the 8-2-1, same with DC power supply test methods.

- DUT working with standalone battery without DC power supply.
- Current/Voltage probe, for battery-powered devices, probing can be done at the battery connectors.
- Controller PC could control and read/record the current/voltage value.
- The beginning condition of battery should be defined as preconditions like 100%, 50%, 20%.
- The test situation should include data throughput, voice and so on.

8.3 Test result analysis

8.3 TBD, will update when starting power consumption test specs definition.

9 References

[1] "GTI Sub-6GHz 5G Device White Paper V2.0" GTI Group, February 2018 http://gtigroup.org/Resources/rep/2018-02-22/11888.html

[2] 3GPP TS37.340 V15.3.0 "Evolved Universal Terrestrial Radio Access (E-UTRA) and NR; Multi-connectivity; Stage 2

[3] Chairman's notes section 7.2.9, RAN WG1 Meeting #94bis, Chengdu, China

[4] 3GPP TS38.304 "User Equipment (UE) procedures in Idle mode and RRC Inactive state"

[5] R1-1810447 "Evaluation Methodology for NR UE Power Saving Study", MediaTek, RAN WG1 Meeting #94bis, Chengdu, China

[6] R1-1810794 "Evaluation methodology for power saving techniques", Intel, RAN WG1 Meeting #94bis, Chengdu, China

[7] R1-1811281 UE Power Saving Evaluation Methodology Qualcomm Incorporated

[8] R1-070674, "LTE physical layer framework for performance verification", Orange, China Mobile, KPN, NTT DoCoMo

[9] R1-1812048 Summary on UE Power Saving evaluation methodology Qualcomm

[10] Chairman's notes section 7.2.9, RAN WG1 Meeting #95, Spokane, USA

[11] 3GPP TS38.331 "Radio Resource Control (RRC) protocol specification"

[12] GTI 5G Sub-6GHz UE Feature List，GTI #23 workshop

[13] TR38.840, "NR; Study on UE Power saving (Release 16)

[14] R1-1901803 Evaluation Results of NR Power Saving Designs, MediaTek RAN WG1 Meeting #96, Athens, Greece