Abstract—With the trends towards Internet of Things (IoT) and massive Machine-Type Communications (mMTC) it is expected that the 5th Generation of mobile communications (5G) will have a significant amount of battery powered devices (e.g. sensors, baggage tags, etc.). Therefore, battery efficiency and duration will be essential, especially for those devices in remote locations and/or restricted areas. It would be difficult to predict all the 5G use cases, for example, that may arise from IoT however it is expected that for some of these the tradeoff between efficient power savings modes and low-latency system access might be essential. In order to solve this tradeoff, called herein User Equipment (UE) sleeping problem, the paper proposes a novel state model for 5G Radio Access Networks (RAN) that relies on a novel state called “connected inactive” where both the UE and the network does not throw away context information. The state is envisioned to be highly configurable in order to address unpredictable use cases possibly with different requirements. It is shown via protocol signaling diagrams that the proposed solution enables a quick and lightweight transition from inactive to active data transmission.

Keywords—5G architecture; UE sleeping; State model; Radio Resource Configuration (RRC) protocol; tight integration; multi-RAT integration;

I. INTRODUCTION

Worldwide initiatives in the mobile industry have started in the past few years in order to define the requirements and principles of what is to be the 5th Generation of mobile communications (5G). In all these initiatives (research projects, industry fora, etc.) it has been acknowledged that in order to fulfill the 5G requirements (e.g. in terms of data rates and latency) the Radio Access Networks (RAN) should be comprised of at least one novel 5G Radio Access Technology (RAT) tightly integrated with the evolution of Long Term Evolution (LTE), to be widely deployed in the 5G time frame [1], [2]. There is also a consensus in the mobile industry that the overall 5G system should support a diverse range of requirements (sometimes divergent) such as the handling of massive Machine-Type Communication (mMTC), low latency (in the order of 1 millisecond) and ultra-reliable MTC (uMTC) and extreme Mobile Broadband (xMBB) services with very high peak data rates (up to 20 Gbps) [3], [4].

Since the first generations of mobile communications, solutions to optimize the power consumption of mobile devices, called herein User Equipment (UE), have been of key importance. Initially this included solutions using periodic Discontinuous Reception (DRX) cycles and paging to support long standby time for voice centric UEs with reasonable sized batteries. In the 3rd Generation (3G) and in the 4th Generation (4G) of mobile communications, with the advent of smartphones, solutions to quickly switch to DRX in-between data bursts have been standardized [5]. With the trends towards Internet of Things (IoT) and mMTC it is expected that in 5G there will be even more battery powered UEs (e.g. sensors, baggage tags, etc.). Therefore, battery efficiency and duration will be essential, especially for those devices which accessibility is limited (e.g. remote locations, restricted areas). Efficient ways to enable the mobile devices to switch their receivers off is required, so that they can run on batteries for a long time. At the same time, the requirement for fast 1st packet transmission from the UEs to the network (or vice versa) is expected to be more stringent in 5G than the previous mobile generations. It is also essential to also reduce the signaling in the network associated with UEs that become active to only send very short data bursts.

Minimizing the latency for 1st packet transmission has the tendency to increase the device battery power consumption since typically the UE would need to perform more preparation actions such as the monitoring of control channels and signal measurements. The problem of creating efficient mechanisms that allows the device to minimize its power consumption and, at the same time, be efficiently reachable by the mobile network and quickly start to transmit data with minimum signaling overhead when it requires is called herein “UE sleeping problem”.

Previous contributions – Recent contributions have addressed the UE sleeping problem. In [6], [7] and references therein the sleeping problem is either addressed by the optimization of existing DRX parameters in LTE or by enhancing the mechanism via new procedures. In [6], a method to reduce the power consumption during DRX “on periods” for MTC devices in LTE is proposed. The method relies on avoiding Physical Downlink Control Channel (PDCCH) decoding, a computationally intensive process that requires substantial processing time that is not useful when the device is not paged. This is achieved thanks to a Quick Sleeping Indication (QSI) independently from PDCCH that allows the device to sleep earlier based on the assumption that MTC devices using LTE are unlikely to be paged every time it wakes up. In [7], the authors have studied the tradeoff between latency and power savings for DRX configured UEs in “Radio Resource Control (RRC) Connected” state and proposed optimizations of the DRX parameters. It has been shown that DRX short cycles are very effective in reducing latency for
active traffic, while shorter inactivity timers are desirable for background traffic to enhance power saving. The authors have also proposed a mechanism to switch between DRX configurations based on different traffic conditions at the UE, using UE assistance (adopted by 3GPP in Release 11). Despite its proven benefit, the solution proposed in [7] is applicable for UEs in “RRC Connected” state where mobility is fully network based (i.e., requiring constant signaling between the UEs and the network). In the 5G time frame billions of devices are expected so that it can be very challenging (and inefficient) to keep them all UEs in “RRC Connected” with full network controlled mobility support otherwise network signaling will explode.

**Our contribution** – This paper addresses the “UE sleeping problem” by proposing a novel state model for the 5G RAN (more specifically, to the RRC protocol). The proposed state model comprises a novel state called “RRC Connected Inactive”, in addition to the existing ones (i.e., “RRC Idle” and “RRC Connected”). The paper presents the characteristics of the novel state and how it improves the overall state model from LTE by [8]:

- Reducing the delay that takes for the UE to access the system and start the data transmission after inactivity periods.
- Reducing the signaling overhead in the radio and the Core Network (CN)/RAN interfaces for UEs coming from inactivity periods that want to transmit small amounts of data (e.g., MTC devices) or short data bursts (e.g., some types of smartphone traffic) after periods where the UE receiver was off.
- Reducing the signaling for moving UEs by enabling an efficient UE-based mobility mechanism during the inactivity periods.
- Enabling a higher level of configurability in order to address a wide range of use cases and services that may have divergent requirements (e.g., in terms of accessability and power consumptions), expected in the 2020 time frame.
- Enabling features such as multi-RAT camping and access where a UE can monitor paging channels of multiple RATs (including the evolution of LTE) and/or simultaneously access them.

These benefits are achieved partially by exploring the principle of “not discarding previously exchanged information” for inactive UEs. In other words, UEs moving to the “RRC Connected Inactive” state will keep parts of the RAN context. This is valid at least for the semi-static information i.e., still valid after inactivity periods such as Access Stratum (AS) security context, UE capability information, etc. In addition to storing the RAN context the UE is allowed to move around within a pre-configured area without notifying the network. This avoids the need in the network to handle the mobility of these devices. A similar principle based on storing the RAN context has been proposed in 3GPP to support highly efficient handling of frequent and infrequent small data transmissions with minimized overhead for system signaling but for LTE “RRC Idle” UEs [16].

**Paper organization** – The paper is organized as follows. Section II describes in more details the “UE sleeping problem”. Lessons learnt from LTE networks are presented in order to motivate the proposed solutions. In Section III, the principles of the novel “RRC Connected Inactive” state are presented. The impact of the novel state to the 5G RAN, also comprising the evolution of LTE is also analyzed. In Section IV, the overall state model for the 5G RAN is presented where some of the benefits highlighted are shown via protocol signaling diagrams.

**II. UE SLEEPING PROBLEM**

In order to enable the UE to save battery it is very important that its receiver (and transmitter) can be switched off when there is no data to be transmitted or expected. This makes possible to achieve significantly longer standby time in the UE compared to “talk” (or active) time. At the same time the UE should be reachable by the network (e.g., via paging) and, if it wants to transmit data, it should quickly be able to access the system. This tradeoff between power savings and quick system access is called herein “UE sleeping problem”.

**A. Discontinuous Reception (DRX)**

In current systems the UE sleeping problem can be controlled via DRX optimizations [5]. A DRX cycle consists of “on periods” during which the UE monitors downlink channels and physical signals (so that it can be reached by the network via paging among other actions) and “sleeping periods” when the UE can switch its receivers off. When the UE is utilizing DRX it probably also uses Discontinuous Transmissions (DTX) however this is not specified in the standards. The tradeoff between power savings and access latency can be controlled by optimizing the settings for the DRX cycles. In the one hand, DRX allows the UE to sleep and, consequently save battery when receivers are off.

**B. DRX applied to the existing states**

Currently, LTE supports two different states in which DRX can be configured: a CN state and a RAN connected state (i.e., where the UE has an ongoing connection) [8]. Some of the characteristics in these two states are the following:

- In the CN state, called Evolved Packet System (EPS) Connection Management (ECM) Idle / RRC Idle, only CN context is stored when the UE is sleeping. UE and network discards the RAN context information when moving to this state. The UE is known at the network only on tracking area level and may move within the cells belonging to the tracking area without informing the network. The UE would, nevertheless, camp in the best cell via cell reselection procedure based on the configuration provided by the network. In this state DRX is used to improve battery performance so the UE is not mandated to constantly monitor paging channels.
and transmit location updates if highly moving. Therefore, DRX mainly impacts the paging performance where longer DRX cycles (for UE power savings) are traded off with the reachability of the UE by the network (that needs to be on to monitor paging channels).

- In the RAN connected state (ECM Connected / RRC Connected) the UE is known on a cell level and mobility is fully network controlled (via handovers). RAN context is present. In this state DRX is used for micro-sleeping periods between bursts of data of better resource utilization. For instance, during a web session short DRX cycles can allow for faster response when data transfer is resumed [7] instead of wasting resources to continuously monitor downlink channels (e.g. PDCCH) while the user is reading a downloaded web page. On the other hand, due to the uncertainties of the traffic demands, DRX can lead to delays when the network wants to reach the UE (e.g. if there is incoming traffic) [10]. In order to reach a good compromise current cellular systems such as LTE supports two configurable DRX cycle, a long and a short with fast switching in between. For uplink traffic the added delay from DRX is less of a problem since the UE can in principle request resources as soon as UL data arrives (assuming there are available scheduling request resources).

C. Lessons learnt from LTE networks

It can be said that out of the two states ECM Idle / RRC Idle is the one used as the primary sleeping state i.e. most of the time the UE sleeps occurs in that state. It has been observed that in LTE networks inactivity timers are typically configured so that inactivity to start transmitting data.

As it can also be observed in the signaling diagram of Fig. 1 there can be many more components in (1) in the case of optional messages and/or error cases such as contention resolution or actions associated to admission control algorithms. In LTE there is a requirement that this delay should be lower than 100 ms and, in LTE-A, lower than 50 ms, so that even lower values should be expected in 5G at least for some services [12]. This scenario is also envisioned in 5G networks with the trends towards IoT [17] where it is expected a large number of devices, each generating a small amount of data.

![Signaling Diagram](image)

**Fig. 1.** Signaling for the “RRC Idle” to “RRC Connected” transition, which is the most typical way to move from inactivity to start transmitting data.

### III. Principles of the Novel Connected Inactive State

As described in Section II, initiating data transmission from “RRC Idle” in LTE involves significantly more signaling (and consequently delays) compared to data transmission from “RRC Connected” with DRX (LTE requires the DRX transition in “RRC Connected” to be lower than 10 ms [12]). On the other hand, it is challenging and not efficient to always keep all UEs in “RRC Connected” state where procedures are optimized for data transmission such as network-controlled mobility, especially assuming that there will be billions of devices in 5G networks.

In order to address the “UE sleeping problem” for 5G devices it is proposed a novel “RRC Connected Inactive” state designed to be used as the primary sleep state for the 5G access. This new state will have the following characteristics:

- **Maintenance of context information** by the UE and the network when the moves from “RRC Connected” to “RRC Connected Inactive”.
- **Widely configurable DRX** cycles (from milliseconds to hours) to support a wide diversity of services with

\[
\text{Transition time} > \text{RA delay} + 3.5 \times \text{radio RTT} + \text{S1 RTT} \quad (1)
\]
different requirements in terms of power consumption and accessibility delays.

- **UE controlled mobility and RAN-based paging** with optimized state transitions for the case where the UE is semi-static i.e. the UE remains in the same location after inactivity timer expires. The concept of camping for Idle UEs is extended to the “RRC Connected Inactive” UEs.

- **Multi-RAT camping and access** where the evolution of LTE is tightly integrated to the 5G RAN.

- **Highly configurable procedures** that may possibly take into account known characteristics at the RAN level such as mobility pattern and traffic characteristics for the different services and performance requirements in terms of delay accessibility.

A. Widely configurable DRX

The wide diversity of 5G use cases will also lead to devices with very different traffic patterns and battery requirements. Since the RRC Connected Inactive is envisioned to be used as the primary sleeping state it is essential to enable a widely configurable DRX cycles in order to comprise the different cases. Some devices may need to sleep for hours and minutes, while others would need to wake up only once a day but still benefit from the fact the RAN context is stored e.g. for network-initiated contact via paging.

B. UE-controlled mobility and RAN-based paging

It is envisioned that in the novel “RRC Connected Inactive” the UE is reached by the network via paging so that the UE can be configured to monitor the paging channel(s). Since this is a RAN state, in order to be reached it is also envisioned that the UE is known at the RAN within a so-called Tracking RAN Area (TRA) and mobility signaling is avoided when the UE moves within a configured TRA or TRA list (possibly similar to an LTE UE in RRC Idle). This would optimize the amount of signaling that could be from moving mMTC UEs.

C. Multi-RAT camping and access

It has been acknowledged that in order to benefit from its widely deployed coverage in the 2020 time frame LTE should be tightly integrated to the 5G RAN [2][3]. A solution relying on a common PDCP and RRC frameworks has been recently proposed [1], inspired in previous research in the area of multi-RAT integration [13]. This will enable the dynamic usage of all available resources (free resources in other access). If this takes too long, those free resources might no longer be there. These scenarios also motivate a common CN connection (e.g. an evolved S1) for both accesses in the case of dual-radio UEs.

A common control plane framework where the evolution of LTE is part of the 5G RAN demands a common state handling for dual radio UEs. Otherwise, any toggling between the novel 5G RAT and LTE coverage due to bad coverage reasons would lead to signaling to update the LTE state. In “RRC Connected Inactive” a dual-radio UE can be configured to either camp on LTE or on a novel 5G RAT (and monitor the respective paging channels accordingly) as in current system. However, one possible alternative would be to configure dual-radio UEs to camp simultaneously on both RATs i.e. monitor paging channels of both RATs and possibly try to access both RATs simultaneously. This could be beneficial for use cases requiring a very fast establishment of multi-connectivity.

D. Highly configurable procedures

In the novel state the network can explore the fact that the RAN context is stored and perform some optimizations for the procedures e.g. based on known characteristics about the services and/or mobility patterns. For instance, by knowing that a given UE is static, the network could maintain the mobility anchor point in a single node so that paging only occurs in a limited area. For certain services, where access is only UE-initiated, paging may not even be configured.

In another example, depending on how predictable the traffic is and the service requirements in terms of access latency the network can configure certain UEs to access the system with some prioritized mechanisms e.g. by providing random access channel configurations dedicated for groups of UEs. Some UEs with services that require ultra-low latency might even be configured to maintain synchronization in downlink and uplink for fast and seamless system access.

Measurement during connected inactive state can be configured depending on the service requirements such as reliability and latency. It is known that transmission diversity significantly improves link level reliability. Thus, a UE with service that requires ultra-high reliability and ultra-low latency might need to monitor the quality of several links during connected inactive state. This enables the UE to quickly resume the connection in several links, for example, for dual-/multi-connectivity. On the contrary, the measurement configuration for UEs with somehow delay tolerant services might be done in an energy efficient approach in a similar fashion as in LTE idle state.

IV. STATE MODEL FOR THE 5G ACCESS

A novel state model is proposed for 5G enabling an efficient UE sleeping, a fast and lightweight transition from sleeping to active states and joint access optimizations. The RAN part of that model is shown in Fig. 2.

---

**Fig. 2.** State model for the 5G architecture (only RAN states)
The model consists of three states: “RRC Idle”, “RRC Connected” and “RRC Connected Inactive”. In the novel model the state transitions from RRC Idle to RRC Connected are expected to occur mainly during the first initial access (e.g. when the UE attaches to the network) or as a fallback case (e.g. when the devices and/or network cannot use the previously stored RAN context). As a consequence, this transition is not expected to occur as often as in LTE.

On the other hand, transitions from “RRC Connected Inactive” to “RRC Connected” are expected to occur quite often and should be optimized as a lightweight and fast transition. Some aspects of this transition are described in the following. Details of this procedure are described in the following.

A. Fast and lightweight transition to RRC Connected

In the novel state model the “RRC Connected Inactive” is proposed to be used as the primary sleeping state. When the UE moves to that state (e.g. via an inactivity timer or via explicit network signaling) both the UE and the network will keep RAN context information that has been obtained during the first RRC connection setup. When the UE attaches to the network and/or moves from “RRC Idle” to “RRC Connected”, examples of context information are the UE capabilities and the security context.

It is envisioned that this state transition is handled by a procedure inspired in what is being defined in [17], called RRC Resume. That procedure involves the resumption of Signaling Radio Bearers (SRB) and Data Radio Bearers (DRB). The connection resume succeeds only if the accessed target node (which can either be LTE or the novel 5G RAT) can find the RAN context and the mobility anchor for the CN/RAN interface. For this reason, a RAN Context ID should be included within a RRC Connection Resume Request that is an SRB0 message [11]. The procedure is triggered by the UE either in response to a paging, when the UE has uplink data in buffer or when it needs to send TRA updates. Upon receiving RRC Connection Resume Request, the network retrieves the UE context based on the RAN Context ID, performs the RAN context fetching and responds with RRC Connection Resume to reconfigure SRBs and DRBs. Upon the reception the UE performs the following actions:

- Re-establish PDCP and RLC for SRBs and DRBs;
- Perform radio resource configuration;
- Perform measurement related actions according to the measurement configuration;
- Resume SRBs and DRBs.

B. Transparent inactivity to the CN

In the proposed state model the CN/RAN context is also stored i.e. the transitions from “RRC Connected” to “RRC Connected Inactive” and the other way around are transparent to the CN. Therefore, incoming packets from the CN may be forwarded to the latest mobility anchor point at the RAN so that transitions from “RRC Connected Inactive” to “RRC Connected” do not involve CN signaling such as the setup of an CN/RAN interface connection as shown in Fig. 1. This is one of the enhancements compared to the solution proposed in [17] which may further reduce the delay in the state transition and the protocol overhead.

C. Delay and protocol overhead analyses

Applying the principle of keeping the RAN context (and its identifier) when the UE moves to RRC Connected Inactive the procedure described in the previous subsection has the following delay budget in its best case scenario (see Fig. 3):

\[ \text{Transition time} > \text{RA delay} + \text{radio RTT} \]

Compared to (1), the transition time shown in (2) represents a reduction of \(2.5 \times \text{RTT} + S1 \times \text{RTT}\). From an RRC procedure perspective, there is approximately 70% signaling reduction in terms of RRC signaling (7 RRC messages vs. 2 RRC messages). One can also notice that the novel procedure does not require the establishment of the CN/RAN connection which may further reduce the protocol overhead and delay. One additional component could be the delay to fetch the RAN context in the network side, which can reasonably assumed negligible in most of the cases where transition will occur without the UE mobility.

When it comes to absolute values a potential difference compared to LTE may exist since the RA delay comprise the delays from lower layer procedures (such as frequency and time synchronization) and random accesses that will likely be different in the novel 5G RAT.
The state transition based on RRC signaling presented in subsection B should be seen as a default option that needs to be supported however, further optimizations may also exist such as the case of MAC control element based signaling for semi-static devices.

Configurable “RRC Connected Inactive” state can be further optimized in case the information on service and/or an application identity is available at network thus differentiating the connectivity for a group of devices sharing the same identity. As an example mMTC group performing infrastructure metering and environmental monitoring can indicate to network with their service identity that the devices are stationary and can be operated with long DRX for very high battery life. As another example uMTC group may demand use-case specific configuration parameters such as extremely high throughput, mobility, critical reliability. In such case the UE service identity can indicate to network that the connection request is coming from an autonomous driving UE and that the service requires ultra-reliable communication with immediate low latency system access to prevent road accidents. Additionally an UE may request for example a service according to its mobility status and/or request the state configuration with the resources dedicated to the service.

V. CONCLUSION

The sleeping problem for 5G devices has been addressed by the proposal of a new state model enabling an efficient UE sleeping, a fast and lightweight transition from sleeping to active states. This new state model contains what we call an “RRC Connected Inactive” state that explores the principle of “not discarding previously exchanged information” for sleeping UEs i.e. UEs in RRC Connected Inactive state will keep parts of the RAN context as recently proposed in 3GPP for LTE [16], [17]. Another novel aspect is that the proposed state model has been designed to efficiently support a tight integration of LTE to the overall 5G RAN.

We can also conclude that the assumption of a common CN and a common CN/RAN connection for both accesses (as discussed in Section III) reduces the amount of network signaling needed between the CN and the RAN for the Connected Inactive to Connected transition. A consequence of this overhead reduction is the reduction of the delay from the sleep to active state.