Small Data Transmission: PHY/MAC

Hyoungsuk Jeon

Ofinno • ofinno.com • July 2021

Small Data Transmission: PHY/MAC

1. Introduction

The IoT (Internet of Things) technologies using machine-type communications are already ordinary in our daily lives. The smart home and office using IFTTT (If This Then That) and OTA (over-the-air) updates in the car are examples that already affect our lives. For some time now, the devices connected to the internet, in use around the world, outnumber the global population. Cisco forecasts that the number of such devices will be more than three times the global population, and M2M (machine-to-machine) connections will account for half of the global connected devices and connections by 2023 as illustrated in Figure 1.

The 3GPP (Third Generation Partnership Project) has developed global standards based on a roadmap toward the most demanding IoT use cases. Since Release 13, LTE (Long-Term Evolution a.k.a. 4G technology) standards have introduced LTE-MTC (Machine-Type Communication) and NB-IoT (NarrowBand-Internet of Things) technologies. The objective of the LTE-MTC and NB-IoT is to reduce device complexity and power consumption and to support extended coverage and high device density. From Release 15, NR (New Radio a.k.a. 5G technology) standards have also introduced technologies such as eMTC (enhanced MTC) and URLLC (Ultra-Reliable Low-Latency Communication) to support connectivity for a massive number of IoT devices with high reliability.

In this whitepaper, we overview 3GPP technologies developed for communications of infrequent and small size data traffic, for example, observed in sensors, smart meters, and wearable devices. This whitepaper categorizes such communications as small data transmission (SDT) and reviews PHY (Physical) and MAC (Medium Access Control) layers procedures developed for the SDT in the 3GPP standards. The current progress and future evolutionary direction of 3GPP IoT technologies will also be discussed here.

2. Small Data Transmission (SDT)

The SDT in the 3GPP simply refers to data transmission in an inactive state[1]. Specifically, the SDT is a transmission for a short data burst in a connectionless state where a device does not need to establish and teardown connections when small amounts of data need to be sent [5].

In the 3GPP standards, the inactive state had not supported data transmission until Release 15. The 3GPP standards basically allowed the data transmission when ciphering and integrity protection are achieved during the connection establishment procedure. Therefore, the data transmission can occur after the successful completion of the establishment procedure between the device and network.

The problem arises as a device stays in the connected state for a short period of time and subsequently releases the connection once the small size data is sent. Generally, the device needs to perform multiple transmissions and receptions of control signals to initiate and maintain the connection with a network. As a payload size of the data is relatively smaller compared with the amounts of the

[1]Throughout this whitepaper, the inactive state refers to a state that an RRC connection is suspended, e.g., the RRC_INACTIVE state in LTE/NR standards and an RRC_IDLE state with a suspended RRC connection in LTE standards [3][4].

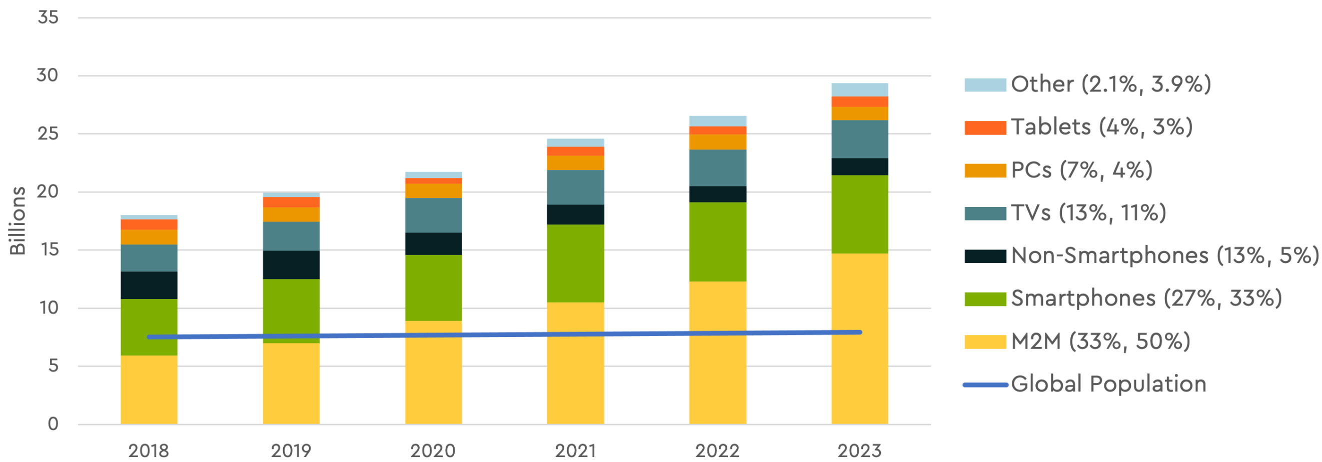

FIG. 1 Projection of number of devices and global population, ([X%, Y%] refers to 2018, 2023 device share) [1][2]

control signals, making a connection for the small data transmission becomes more of a concern for both the network and the device due to the control signaling overhead.

The 3GPP has developed the SDT procedure to enable data transmission in the inactive state over the existing LTE and NR standards. The device initiates the SDT procedure by transmitting an RRC request message (e.g., SDT request message) and data in parallel instead of transmitting the data after the RRC request message processed by a network. Additional transmission and/or reception are optional. The device performs this SDT procedure without transition to the connected state (i.e., without making a connection to the network).

The SDT enables for the network to accept data transmission without signaling intensive bearer establishment and authentication procedure required for the RRC connection establishment or resume procedure. For example, in the SDT procedure, the device needs only one immediate transmission of a transport block (TB) that contains data and RRC request message. Furthermore, the device does not need to perform procedures (e.g., radio link monitoring) defined in the connected state since the RRC state is kept as the inactive state. This results in improving the battery life of the device by avoiding control signaling unnecessary for transmission of small size data [5].

The principle of the SDT is very simple. The network configures radio resources beforehand for the data transmission in the inactive state. For example, if the conditions to use the configured radio resources satisfy, the device transmits data and the RRC request message together via the configured radio resources. In the 3GPP standards, there are two types of the SDT depending on the ways to configure the radio resources: (1) SDT using a random access (RA) and (2) SDT using preconfigured radio resources.

Figure 2 illustrates different types of the SDT referred in 3GPP LTE and NR standards. The SDT using the random access in LTE and NR standards is referred to as an EDT (early data transmission) and RA-SDT (Random Access based SDT), respectively. For both the EDT and the RA-SDT, the device performs data transmission using shared radio resources of the random access procedure. Thus, the contention with other devices can occur over the access to the shared radio resources. The shared radio resources for the SDT are broadcast by system information and are configured as isolated from the one for a non-SDT RA procedure, i.e., the legacy RA procedure. On the other hands, the CG-SDT uses the preconfigured radio resources dedicated to the device. The SDT using the preconfigured radio resource is referred to as transmission via PUR (Preconfigured Uplink Resource) in the LTE standards. The NR standards refers the SDT using the preconfigured radio resource as CG-SDT (Configured Grant based SDT). The network configures the configuration parameters of

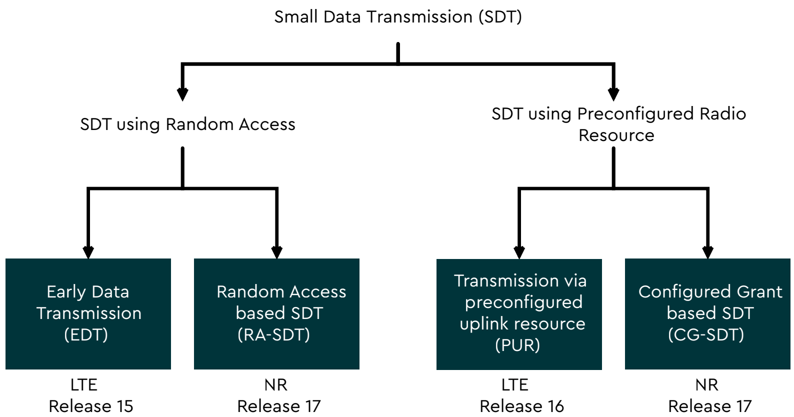

FIG. 2 Different types of SDT

the preconfigured radio resources when transiting the device in the connected state to the inactive state. For example, an RRC release message transmitted from the network for a connection release contains the configuration parameters of PUR or CG-SDT. No contention is expected for the SDT using the preconfigured radio resource since the configuration parameters are dedicated to the device.

3. LTE standards for SDT

The SDT in the LTE standards is developed for the LTE-MTC and NB-IoT devices. Therefore, the SDT procedure in the LTE standards has been optimized based on the capability of the LTE-MTC and NB-IoT devices. The limited PDCCH (Physical Downlink Control Channel) monitoring capability and use cases of the LTE-MTC and NB-IoT applications determined the SDT procedure in detail during the 3GPP discussion. For example, the LTE standards do not expect additional multiple transmissions subsequent to the SDT according to the use case analysis of the LTE-MTC and NB-IoT applications [6].

In Release 15, the 3GPP introduced the MO (Mobile-Originated) EDT as the first SDT mechanism in the LTE standards for uplink data transmission in the inactive state. Since then, the LTE standards for the SDT have evolved, in Release 16, in a way to support MT (Mobile-Terminated) EDT for downlink data transmission and to support data transmission via the PUR. In Release 17, it is expected that EDT and transmission using PUR will be further enhanced for the uplink and downlink transmission efficiency as well as the power consumption of LTE-MTC and NB-IoT devices.

3.1 EDT (Early Data Transmission)

In LTE standards, an RRC layer of the device checks conditions for initiating the EDT. The following are example conditions that allow the device to initiate the EDT [3]:

- System information received from a network (e.g., Evolved Packet Core or 5G core) where the device connects should include EDT parameters and an indication of supporting the EDT
- The device should support the EDT and have a stored value valid to determine security keys for the EDT
- An establishment cause should be suitable for the EDT
- The data size should be less than or equal to a TB size indicated by the system information.

Once the conditions are fulfilled, the RRC layer configures lower layers (e.g., MAC and PHY layers) to use the EDT for uplink or downlink data transmission. Accordingly, the MAC layer of the device initiates an RA procedure for the EDT and the PHY layer also starts transmission and reception.

The LTE standard uses a four-step contention based RA procedure for the EDT as illustrated in Figure 3.

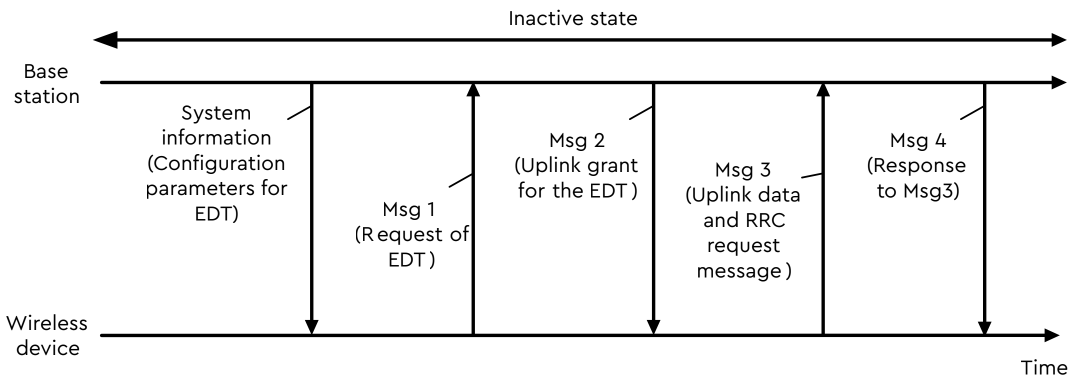

FIG. 3 RA procedure for the EDT

Therefore, the RA procedure for the EDT consists of, as the non-SDT RA procedure, a preamble transmission (Msg1), RAR (random access response) reception (Msg2), and a contention resolution using Msg3 transmission and Msg4 reception.

The configuration parameters of the RA procedure for the EDT are broadcast by system information of the cell. Many of the configuration parameters are shared with the non-SDT RA procedure, but some of them are configured separately from the one for the non-SDT RA procedure. For example, the network configures the resource of the preamble transmission for the EDT such that it is isolated from the one for the non-SDT RA procedure. Thereby, the network can identify whether the device requests the EDT, depending on where the preamble is received.

For the LTE-MTC device, the configuration parameters of the RA procedure for the EDT can be configured per a coverage level. For the NB-IoT device, the configuration parameters can be configured in both anchor carrier and non-anchor carrier. Once the MAC layer initiates the RA procedure for the EDT, the MAC layer can cancel the initiated EDT and inform of the cancellation to the upper layers. For example, if a coverage level changed during the RA procedure does not support the EDT or if the data size is larger than a TB size that the changed coverage level supports, the MAC layer can cancel the initiated EDT. Also, the network can use the RAR to indicate a fallback to the non-SDT RA procedure from the RA procedure initiated for the EDT. A reserved bit of a non-SDT MAC RAR format is used as the indication of the fallback.

More specifically, the RA procedure initiated for the EDT is as follows:

- Preamble (Msg1) transmission
    - The configuration parameters for the preamble transmission can be configured separately from the one for the non-SDT RA procedure. For example, the system information can indicate: (1) PRACH (Physical Random Access Channel) configuration configured for the EDT; and/or (2) a list of preambles reserved for the EDT in the cell. Therefore, the network that receives the preamble associated with the EDT can identify that the device requests the EDT and accordingly provide a proper size of an uplink grant for the EDT using the RA procedure.
- RAR (Msg2) reception
    - Once the device transmits a preamble for the EDT, the device monitors PDCCH using a RA-RNTI (Random Access-Radio Network Temporary Identifier) for a RAR corresponding to the preamble. The way to calculate the RA-RNTI for the EDT is the same as the legacy LTE-MTC and NB-IoT. However, no conflict with the legacy LTE-MTC and NB-IoT is expected since the dedicated PRACH

configuration and/or a preamble for the EDT results in isolating the RAR for EDT from the one for non-EDT.

  - The MAC RAR that the device receives corresponding to the preamble for the EDT includes a field indicating that an uplink grant in the RAR is for EDT. If the value of the field indicates that the uplink grant is not for the EDT, the device cancels the initiated EDT but continues the subsequent transmission of the initiated RA procedure with a newly constructed TB.

- Contention resolution (Msg3 and Msg4)

  - The actual user data is carried in the TB of the Msg3 transmission. The user data and an RRC message requesting the EDT are multiplexed into the TB and transmitted via the uplink grant provided by the RAR. The uplink grant of the RAR also indicates a repetition level for the transmission of the TB. In the EDT, retransmission of the TB using HARQ (Hybrid Automatic Repeat reQuest) is also supported.

  - The contention resolution for the EDT follows the non-SDT RA procedure. The Msg4 received as a response to the TB includes a contention resolution identity and a response to the request of the EDT. The response can indicate the completion of the EDT, rejection of the EDT, or transition to the connected state.

3.2 PUR (Preconfigured Uplink Resource)

The LTE standard introduces a new RRC message, a PUR configuration request message, that contains a traffic profile of the device for the transmission using the PUR [3]. The procedure of the PUR configuration request message occurs when the device is in the connected state. The procedure comprises a single one-way transmission from the device to the network. No response from the network to the device is defined. The purpose of the procedure of the PUR configuration request message is to indicate that the device is interested to be configured with PUR and provide PUR related information to the network, or that the device is no longer interested to be configured with PUR [3]. For example, the PUR configuration request message can indicate a requested number of PUR occasions, a requested periodicity between consecutive PUR occasions, and a requested size of TB for the PUR occasions. Furthermore, the device can also indicate if RRC response message is preferred by the device for acknowledging the reception of a transmission using PUR [3]. The network can determine configuration parameters of the PUR based on the traffic profile that the device sent via the PUR configuration request message. The network sends, to the device, the determined configuration parameters together with an RRC connection suspend indication in the RRC release message.

Once the device transits to the inactive state, the device keeps the configuration parameters of the PUR. The configuration parameters are only valid on a cell where the device receives the RRC release message that contains the configuration parameters of the PUR. If the device moves to the other cell, the device can still keep the configuration parameters but cannot use them for the other cell. The device releases the configuration parameters when the device performs an RRC connection establishment or resume procedure on the other cell.

The initiation of the transmission using PUR is similar to that of the EDT. The RRC layer of the device also checks whether the transmission using the PUR can be initiated. More specifically, the RRC layer can initiate the transmission using the PUR when the following conditions are fulfilled [3]:

- The device should have a valid PUR configuration for the current cell

- The device should have a valid timing alignment value for transmission using PUR

- The device should have a stored value valid to determining security keys

- An establishment cause should be suitable for the transmission using PUR
- The data size should be less than or equal to a TB size indicated by the system information.

Figure 4 illustrates PUR configuration and release. The device receives an RRC release message that contains the configuration parameters of the PUR from the network. The device can receive the RRC release message for transition from the connected state to the inactive state or as a response to the transmission using PUR in the inactive state. The RRC release message contains the connection suspension indication with the configuration parameters of the PUR. The configuration parameters of the PUR can be for a single PUR occasion or for multiple PUR occasions that periodically occur. The periodicity of the PUR occasions is determined based on use cases of the LTE-MTC and NB-IoT. The value of periodicity is in the unit of H-SFN (Hyper System Frame Number) duration (10.24 seconds) in the LTE standards. In Release 16, the smallest value of the periodicity is 8 H-SFN (about 82 seconds) and the largest value of the periodicity is 8192 H-SFN (about 24 hours). In LTE standards, the 3GPP introduces an implicit release of PUR configuration for the case that the device has no activity for a while. The configuration parameters of the PUR indicate the release timing of the PUR configuration using a number of consecutive PUR occasions skipped. The device can release the PUR configuration after the number of consecutive PUR occasions have been skipped.

The device determines that the timing alignment value for transmission using PUR is valid if a time alignment timer is running and/or if a measured RSRP (Reference Signal Received Power) value has not been changed more than threshold(s) (e.g., increase threshold and/or decrease threshold) since the last transmission using the PUR. The validation using the time alignment timer and the measured RSRP value(s) are optional. Thus, the device uses such a validation method only when the respective configuration parameter(s) are configured. The value of the time alignment timer needs to be long enough to guaranteed at least two consecutive PUR occasions. Thus, the value of the time alignment timer for the PUR is defined as a multiple of the periodicity of the PUR occasions. The timing alignment value can be adjustable based on a timing advance command received via MAC CE (Control Element) or timing advance adjustment value received via DCI (Downlink Control Information).

4. SDT development in NR standards

In 2020, the 3GPP started the WI (Working Item) of the SDT for Release in 17 NR standards [5]. The SDT in NR standards is being built on two-step RA procedure, four-step RA procedure, and Configured Grant (CG) Type-1 that have already been specified as part of Release 15 and Release 16 in NR standards.

As the NR standards support the massive number of IoT devices, it is expected that the SDT in the NR standards can guarantee KPIs (Key Performance Indicators, e.g., low latency, high reliability,

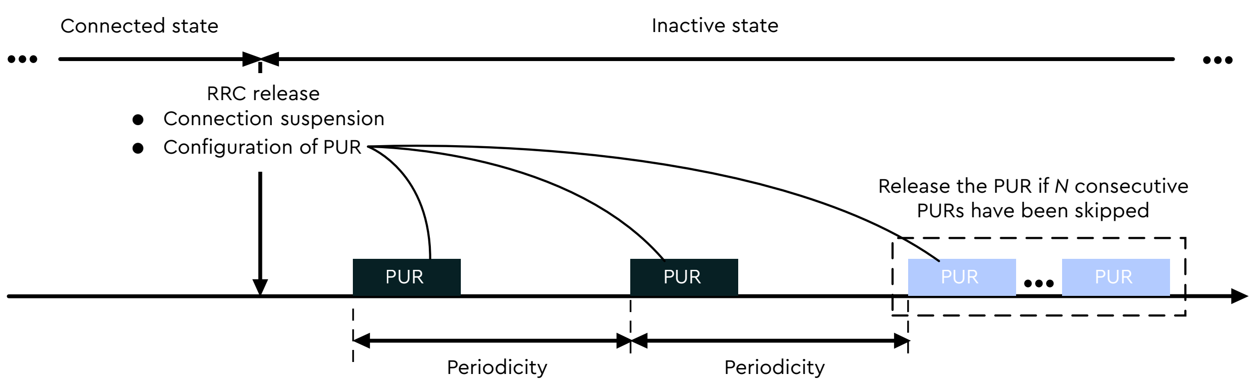

FIG. 4 PUR configuration and release

positioning accuracy) specified for Industrial, Enterprise, and Consumer use of IoT devices. The example SDT use cases in the NR standards are broader than that of the LTE standards, e.g., from the smartphone applications to specialized devices such as wearables, sensors, and/or smart meters.

The 3GPP is currently developing the SDT for NR standards based on the EDT and the PUR. Many features introduced in LTE standards will be reused as the baseline of the SDT design in the NR standards. From the lower layer's perspective, the main challenge in the SDT for NR standards is to introduce NR features that have been designed/optimized for operation in the connected state as an operation in the inactive state. For example, the NR features include the beam-centric transmission and management, flexible frame structure and scheduling (e.g., numerology, control resource set, bandwidth part switching, supplementary uplink), higher frequency operations, and so on. Some of those NR features may be a burden for the device and network to operate in the inactive state, where more energy efficiency and higher resource efficiency are required.

Currently, the 3GPP is having active discussion in the SDT design for NR standards and already made some agreements for high level operations of the SDT in NR standards. The RA-SDT in the NR standards supports the four-step RA procedure as well as the two-step RA procedure consisting of Msg A transmission and Msg B reception. The RA resource for the SDT in the NR standards will also be configured as isolated from the one for the non-SDT RA procedure. The CG-SDT in the NR standards is based on a Configured Grant Type 1 that is initiated and suspended by an RRC message. As LTE standards, the configuration parameters of the CG-SDT will be provided by an RRC release message. Multiple configurations of the CG-SDT are also agreed, and the CG-SDT can be configured in both a NUL (Normal Uplink) carrier and a SUL (Supplementary Uplink) carrier.

Main differences between the LTE and NR standards are highlighted based on the 3GPP agreements in the following subsections.

4.1 Subsequent transmissions

One of main differences from the LTE standards is that the SDT procedure in the NR standards allows multiple transmissions subsequent to an initial transmission of the SDT. Figure 5 illustrates the SDT procedure with multiple subsequent transmissions in the NR standards. Based on the 3GPP agreements, the SDT starts with transmitting an RRC request message and data together as an initial transmission of the SDT and can continue one or more subsequent transmissions. The device can perform the initial transmission using the RA-SDT or CG-SDT. It is expected that the device transmits a BSR (Buffer Status Reporting) or any type of assistance information in the initial transmission to request the one or more subsequent transmissions. The SDT session comprising the initial transmission and the one or more subsequent transmissions ends when the device receives an RRC release message. It is expected that the device needs to perform PDCCH monitoring continuously (e.g., non-DRX mode in the connected state) or discontinuously (e.g., DRX mode in the connected state) during this SDT session to receive a dynamic grant and/or feedback (e.g., ACK/NACK) from the network.

The multiple transmissions occurring during the SDT procedure may look like a communication session of the connected state temporarily and virtually created over the inactive state as illustrated in Figure 5. Indeed, the 3GPP already made some agreements for the SDT based on the device behavior in the connected state. For example, the device starts the PDCCH monitoring for the subsequent transmission(s) with C-RNTI after RA-SDT as it does after completion of the non-SDT RA procedure in the connected state. For the CG-SDT, multiple HARQ processes and retransmission by a dynamic grant are agreed.

The 3GPP is discussing how to maintain the link for the multiple subsequent transmissions during the inactive state. For the MAC layer, the SR (Scheduling Request) resource will not be configured in Release 17. Thus, it is expected that the BSR triggered during the inactive state will trigger the RA procedure due to the SR resource unavailable in the inactive state.

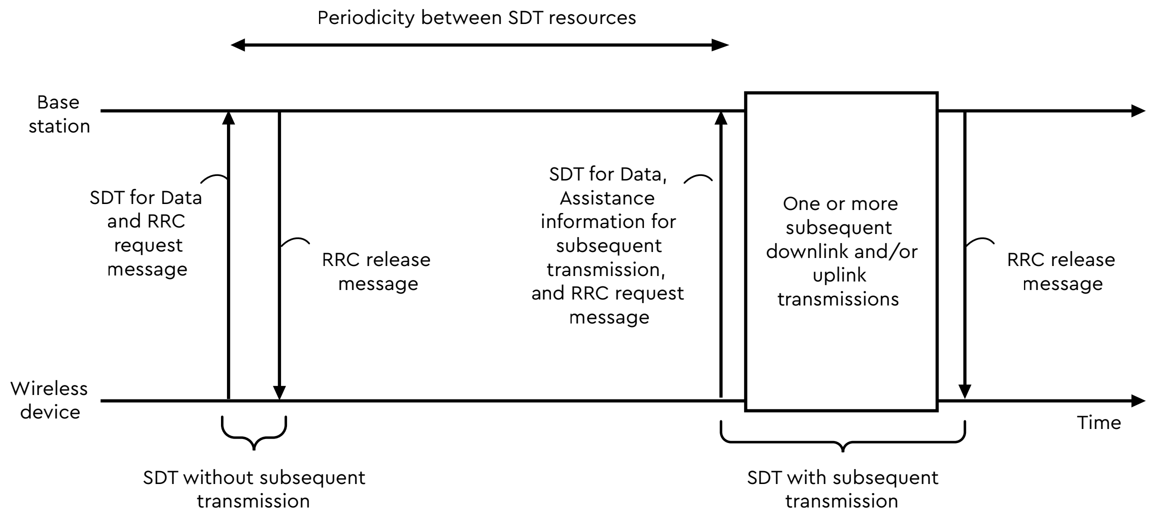

FIG. 5 Subsequent transmission(s) of SDT

The PHR (Power Headroom Reporting) functionality is supported, e.g., for the subsequent transmissions of the SDT. The beam failure detection and recovery procedure and the unlicensed band operation will be discussed once the basic functionalities of the SDT operation specified more in detail. The 3GPP RAN1 working group recently began discussion about the PHY layer design for the SDT. It is expected to discuss the closed-loop power control, beam management, and measurement reporting for the subsequent transmissions.

4.2 SDT type selection

In the LTE standards, a selection between the EDT and transmission using PUR is left up to the device implementation. However, the NR standards agree a selection hierarchy among the RA-SDT based on two-step RA procedure (referred to as two-step RA-SDT), the RA-SDT based on four-step RA procedure (referred to as four-step RA-SDT), and the CG-SDT. Based on the agreements, the device first performs the uplink carrier selection between the NUL carrier and SUL carrier. On the selected uplink carrier, the device selects the CG-SDT as the SDT in the inactive state if the resources for the CG-SDT are configured on the selected uplink carrier and are valid. If the resources for the CG-SDT are unavailable or invalid, the RA-SDT or the non-SDT RA procedure will be selected. For example, if the resources for the two-step RA-SDT are configured on the selected carrier and criteria to select two-step RA-SDT is met, then two-step RA-SDT is chosen. Otherwise, the device selects the four-step RA-SDT if the resources of four-step RA-SDT are configured on the selected uplink carrier and criteria to select four-step RA-SDT is met. If none of the resources for CG-SDT and RA-SDT is available or valid, the device establishes a connection with the network based on initiating the non-SDT RA procedure.

4.3 CG-SDT configuration release

The configuration release mechanism between the PUR in LTE standards and CG-SDT in NR standards is different. In NR standards, the implicit release introduced in the LTE standards will not be supported for the CG-SDT. The main reason is due to the high complexity as well as the possible misalignment between the device and network. Instead, the 3GPP agrees that the device releases the configuration of the CG-SDT when the timing alignment timer expires in the inactive state. This is also different from the LTE standards where the device keeps the configuration of the PUR when the timing alignment timer expires.

5. Future evolution direction of 3GPP IoT technologies

As earlier mentioned, we have seen the rapid increase of IoT devices around our daily lives—in our homes, in our communities, and in the office as well as from the ground up to the sky. More flexible radio technologies that can provide larger bandwidth, more efficient use of radio resources, less power consumption, and wider coverage are needed to support the massive connections of IoT devices in the near future.

The 3GPP standards for IoT applications will continue to evolve, as the IoT market demands more specialized requirements for diverse applications. The SDT is in a clear position to lead the IoT business in success as the need for a simple communication mechanism is addressed for the IoT device that has intermittent small data packets in the inactive state. In the LTE standards, many efficient resource control mechanisms have been implemented for such a small data packet transmission in the inactive state. In NR standards, the SDT development just started along the path paved by the LTE standards.

After Release 17, enhancements in the NR standards are expected in diverse areas. Enhancements to unlicensed and shared bands and to the non-terrestrial network are expected to allow the industry and private enterprise to accommodate various types of connection needs. The integrated design with the broadcast/multicast IoT applications in massive IoT use case are also expected.

6. Reference

1. *Cisco Annual Internet Report (2018–2023)* White paper, March 9, 2020, Available from: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html.
2. *United States Census Bureau*, https://www.census.gov.
3. 3GPP TS 36.331 v16.4.0, Evolved Universal Terrestrial Radio Access (E-UTRA) Radio Resource Control (RRC) Protocol specification (Release 16), March 2021.
4. 3GPP TS 38.331 v16.4.1, NR Radio Resource Control (RRC) Protocol specification (Release 16), March 2021.
5. 3GPP RP-193252, Work Item on NR small data transmission in INACTIVE state.
6. 3GPP TR 22.891 v14.2.0, Feasibility Study on New Services and Markets Technology Enablers; Stage 1 (Release 14), September 2016.

About the Author:

Hyoungsuk is an inventor working primarily on 3GPP air interface, focusing on PHY/MAC layers procedures. Since he joined Ofinno, he has been involved in more than one hundred patent drafting and prosecutions as a lead inventor and a co-inventor. Prior to joining Ofinno, Hyoungsuk developed self-organizing network (SON) automation systems that were commercially tested and deployed in live networks of major operators in various countries. As a researcher, Hyoungsuk has authored over forty journal and conference papers. Two of his journal papers were nominated in IEEE's Best Readings in Communications and Information Systems Security. Hyoungsuk earned his B.S. in Electrical Engineering from Dongguk University and his M.S. and Ph.D. in Information and Communication Engineering from Korea Advanced Institute of Science and Technology.

About Ofinno:

Ofinno, LLC, is a research and development lab based in Northern Virginia, that specializes in inventing and patenting future technologies. Ofinno's research involves video and communication technologies, including 5G and 6G Radio and Core networks, video compression technologies and transport. Ofinno's inventions have an impressive utilization rate and have been adopted by the standards at the center of advancement of these technologies. Our innovators not only create the technologies, they oversee the entire process from design to the time the technology is sold. For more information about Ofinno, please visit www.ofinno.com.