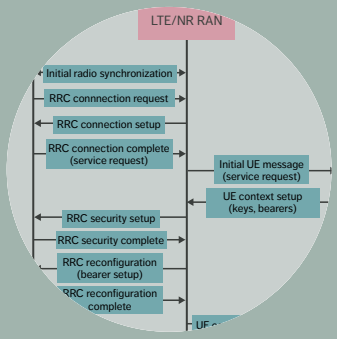
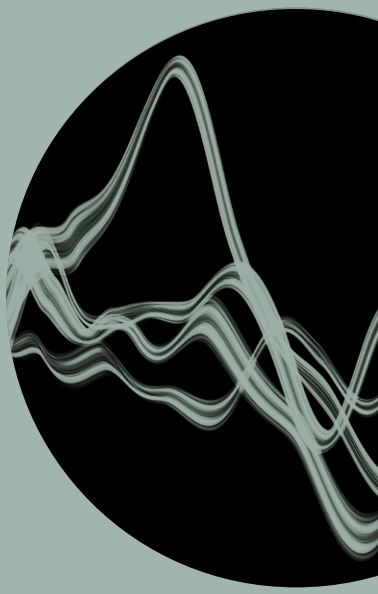
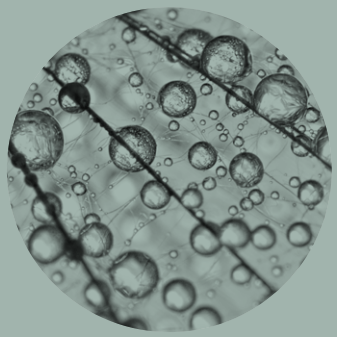
CHARTING THE FUTURE OF INNOVATION | #06·2019

ERICSSON TECHNOLOGY Review





THE INACTIVE STATE IN 5G NEW RADIO



ERICSSON

Meeting 5G latency requirements

WITH INACTIVE STATE

Reducing the amount of signaling that occurs during state transitions makes it possible to significantly lower both latency and battery consumption – critical requirements for many Internet of Things and 5G use cases, including enhanced mobile broadband.

ICARO LEONARDO DA SILVA, GUNNAR MILDH, PAUL SCHLIWA-BERTLING, MAGNUS STATTIN, ALEXANDER VESELY

Many of the performance improvements in 5G New Radio (NR) that are designed to support new Internet of Things (IoT) use cases such as critical control of remote devices and smart transport [1] are based on lessons learned from research and development on 4G LTE networks. One example of this relates to the transition of wireless devices from a power-saving state where data is not exchanged (idle state) to a connected state optimized for data transmissions (connected state).

■ Studies show that a wireless device's transition from a power-saving (idle) state to a connected state is the most frequent high-layer signaling event in existing 4G LTE networks, occurring about 500-1,000 times a day. The transition comprises an extensive signaling sequence between the device and the network, and between network nodes, which can lead to consumer latency issues and high battery consumption.

The combination of 4G/5G research activities and lessons learned from legacy networks has made it possible to develop solutions that reduce the amount of signaling required at these transitions, thereby lowering both latency and battery consumption significantly. The decreased signaling in the network also results in an increase in overall system capacity.

Ericsson's contributions to the 3GPP standardization of solutions in this area include a new Radio Resource Control (RRC) state model adopted in the standalone version of the 5G NR

standard. Improved connection, state and mobility handling are key elements of efficient support for current and future 5G use cases with a large and growing number of devices.

Concept development of the inactive state

Allowing wireless devices to enter a low-power state when they are not transmitting or receiving data has always been an important part of achieving a balance between good communication performance and acceptable battery consumption. For many years, two states – connected and idle – were sufficient to meet most needs.

The development of the inactive state has largely been driven by the growing field of Machine-type Communication (MTC). In most MTC scenarios, the amount of data that wireless devices typically exchange with the network is small and usually not urgent enough to justify the high battery consumption required to handle all the signaling involved in the legacy idle-to-connected transition. To address this issue, Ericsson played a leading role in developing the transition enhancements that were introduced in 4G LTE Rel-13, in which two new procedures were standardized: suspend and resume.

In the suspend procedure, the user equipment (UE) – the 3GPP name for wireless devices –stores its radio configuration and security parameters when it transitions from connected to idle. Then, when it needs to connect to the network again (due to some uplink (UL) data being available to transmit, for example) the UE triggers the resume procedure. This involves restoring the previously stored configuration and resuming the connection without the need for extensive signaling with the core network (CN) or having to reestablish security, for example. The resume procedure is similar to the sleep state of a computer, which enables work to be paused and resumed later without repeating tedious start-up procedures.

In parallel with the 4G LTE work that was completed in 2015, Ericsson was also working on the 5G concept, which included challenging latency requirements and providing support for a variety of new and emerging use cases. Without the constraint to comply with an existing state model, it was possible to further optimize the suspend/resume solution in 5G NR by introducing a new state known as inactive. The key benefits of the inactive state are that it significantly reduces latency and minimizes the battery consumption of both smartphones and MTC devices.

In the latter half of 2015, we began to promote the inactive state externally in the context of the Ericsson-led 5G-PPP European project METIS-II, the main 5G pre-standards project [2]. The main goal of the project was to facilitate research discussions with industry players (UE vendors, network vendors, network operators, academic partners and so on) about technical components to bring to the 3GPP during the 5G standardization work.

Key terms

Connected state – The UE is actively involved in sending or receiving data or signaling. Mobility is controlled by the RAN.

Idle state – The UE is in a power-saving state and is known at tracking-area level in the CN.

Inactive state – The UE is in a power-saving state and is known on RNA level in the RAN. Transition to the connected state is optimized.

System functions ↓	UE RAN states → Idle	Connected	Rel-13 suspend	5G inactive
Mobility management	CN	RAN	CN	RAN
Paging trigger	CN	n/a	CN	RAN
UE configuration data storage	CN	CN and RAN	CN and RAN	CN and RAN
UP contexts in RAN	No	Yes	Yes, but DL packets not sent to RAN	Yes, DL packets sent to RAN

Figure 1 Comparison of the allocation of system functions

Figure 1 shows the allocation of basic system functions in diverse UE states, highlighting the evolution from Rel-13 suspend to 5G inactive.

5G NR inactive state procedures

In 2016, it was agreed that the inactive state would be introduced in 5G NR [3], and the specifications were finalized and approved in December 2018 [4, 5]. The most notable enhancements are the suspend and resume procedures, as well as RAN-based location management and RAN paging for UEs in the inactive state. In the suspend procedure, both the UE and the RAN store information about the UE transition from connected to inactive, along with the UE radio protocol configuration. The resume procedure optimizes the transition from inactive to connected by restoring the UE radio protocol configuration. RAN-based location management and RAN paging make it possible for UEs in the inactive state to move around in an area without notifying the network.

Suspend

The main principle of the inactive state is that the UE is able to return to the connected state as quickly and efficiently as possible. When the UE transitions to inactive, both the UE and the RAN store all the information necessary to quickly resume the connection. The message that transitions the UE to inactive state contains a set of parameters used for inactive state operation, such as a RAN Notification Area (RNA) within which the UE is allowed to move without notifying the network. Further, it includes parameters used for secure transition back to the connected state, such as a UE identifier and security information needed to support encrypted resume messages.

Resume

An inactive UE may initiate a resume procedure when there is a need to transmit data or signaling, for example. In this case, the UE transmits an RRC resume request that includes the UE identifier (provided by the serving node to identify the UE's configuration repository) and a security token to verify the legitimacy of the resume request.

Studies of 4G networks show that, in most cases, UEs that leave the power-saving state return to the same RAN node they were previously served by. If, however, the UE resumes in a cell served by a different RAN node, that target node will retrieve the UE configuration from the serving node based on the UE identifier.

After the UE configuration is successfully retrieved, the target node resumes the stored configuration at the UE and applies any necessary modifications, such as the configuration of measurements and the addition or removal of bearers. The respective RRC resume message is integrity protected and encrypted using the security context stored in the network and the UE.

As illustrated on the right side of *Figure 2,* the resume procedure reduces the number of RRC messages exchanged over the radio interface between the UE and the RAN to three (down from seven for idle state). RRC resume also has the possibility of using efficient delta signaling – in which only changed parameters are signaled – to restore the configuration of a UE in the inactive state. This option is not possible for UEs in the idle state.

The reduction in RRC signaling significantly lowers the access latency experienced by UEs, which leads to more responsive end-user service and the ability to support new use cases. It also reduces the power consumption for devices such as battery-powered sensors that only send small, infrequent reports (often less than 100 bytes of data).

RAN-based location management and RAN paging

The transition from the connected to the inactive state is designed to be invisible to the 5G CN. As a result, even when the UE is in the inactive state, the 5G CN treats it as though it were in the connected state – that is, the UE-associated signaling and user-data connection between the 5G CN and the RAN continues. Mobile-terminated signaling and user-plane (UP) data is sent from the CN to the RAN node currently serving the UE.

When the serving RAN node receives signaling or data for a UE in the inactive state, it initiates RAN paging. The paging is performed in an RNA that consists of one or more cells and was assigned to the

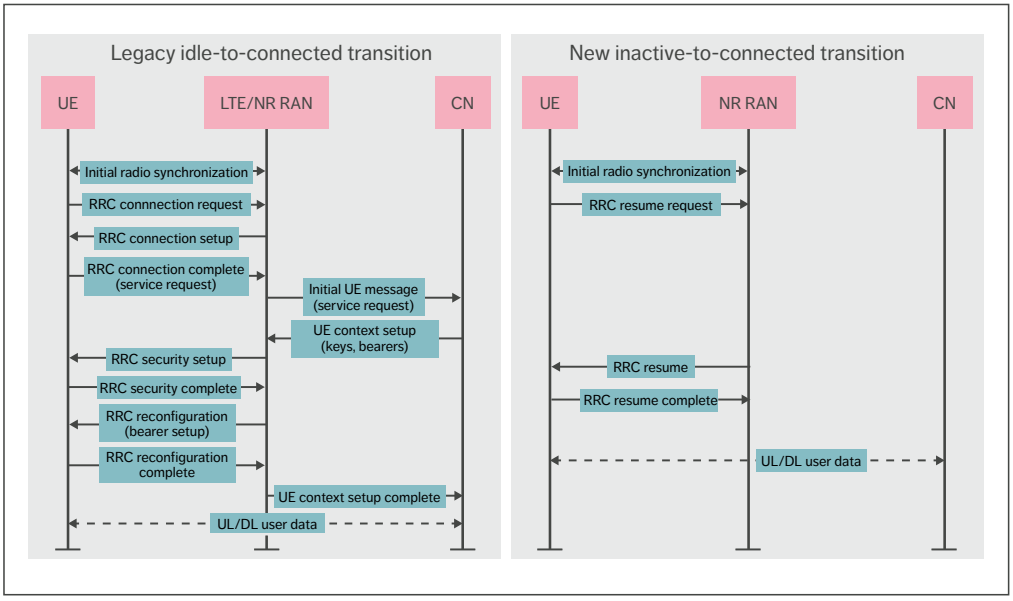


Figure 2 Comparison of signaling involved in legacy idle-to-connected transition (left) versus inactive-to-connected transition (right)

UE when it was ordered to enter the inactive state. When a UE in the inactive state moves to a cell that is not part of its currently assigned RNA, the UE performs a location-update procedure that enables the RAN to update the assigned RNA to the UE.

As in earlier cellular systems, there is a trade-off between the paging load and the amount of location-update signaling. Larger paging areas have more paging load but less location update signaling than smaller paging areas.

Key technology aspects

The most notable technology aspects within the NR inactive state concept adopted by 3GPP are support for encrypted response messages, smart RAN paging, RAN architecture support and fallback to legacy procedure.

Support for encrypted response messages

One of the main components driven by Ericsson in the NR inactive state concept adopted in 3GPP is the ability to encrypt the response message (resume or suspend/release) from the network. This differs from the 4G LTE resume concept adopted in Rel-13, where this message is integrity protected, but sent unencrypted.

To enable the encryption capability, the 3GPP adopted a solution proposed by Ericsson in which the network provides the UE with a security parameter in the release message to the inactive state. The UE uses the security parameter to calculate a new security key to be used when it resumes.

The ability to encrypt the resume response message in 5G NR is advantageous because it makes it possible to use a single, secure message to:

- Reconfigure any parameter in the UE when transitioning to the connected state.
- Release the UE to the idle state (the release message could also include redirection information to another frequency or radio access technology that could be used for voice fallback to LTE, for example).
- Resuspend the UE to the inactive state when it is performing a location-update procedure, for example, so that it only consumes two messages in total (request and response).

Smart RAN paging

For any UE in the connected state, the RAN node receives paging assistance information related to potential paging triggers, such as QoS flows or signaling, from the 5G CN. This information, in combination with other information that the RAN has about the UE, can be used by the RAN node to select and apply a smart paging strategy that is aligned with the characteristics and requirements of the UE and paging-triggering services.

For example, the RAN can configure UE-specific RNAs that make it possible to reduce the total signaling load by configuring small RNAs for stationary UEs (optimized for low paging load) and larger RNAs for moving UEs (optimized for low location update signaling load).

Terms and abbreviations

AMF – Access and Mobility Function | **CA** – Carrier Aggregation | **CN** – Core Network | **CP** – Control Plane | **CU** – Central Unit | **DC** – Dual Connectivity | **DL** – Downlink | **DRX** – Discontinuous Reception | **DU** – Distributed Unit | **EDT** – Early Data Transmission | **E-UTRA** – Evolved Universal Terrestrial Radio Access | **GNB** – GNodeB | **IoT** – Internet of Things | **MAC** – Medium Access Control (protocol) | **MTC** – Machine-type Communication | **NB-IoT** – Narrowband Internet of Things | **NG-RAN** – Next-Generation RAN| **NR** – New Radio | **PPP** – Public-private Partnership | **RNA** – RAN Notification Area | **RRC** – Radio Resource Control | **UE** – User Equipment | **UL** – Uplink | **UP** – User Plane | **UPF** – User-plane Function

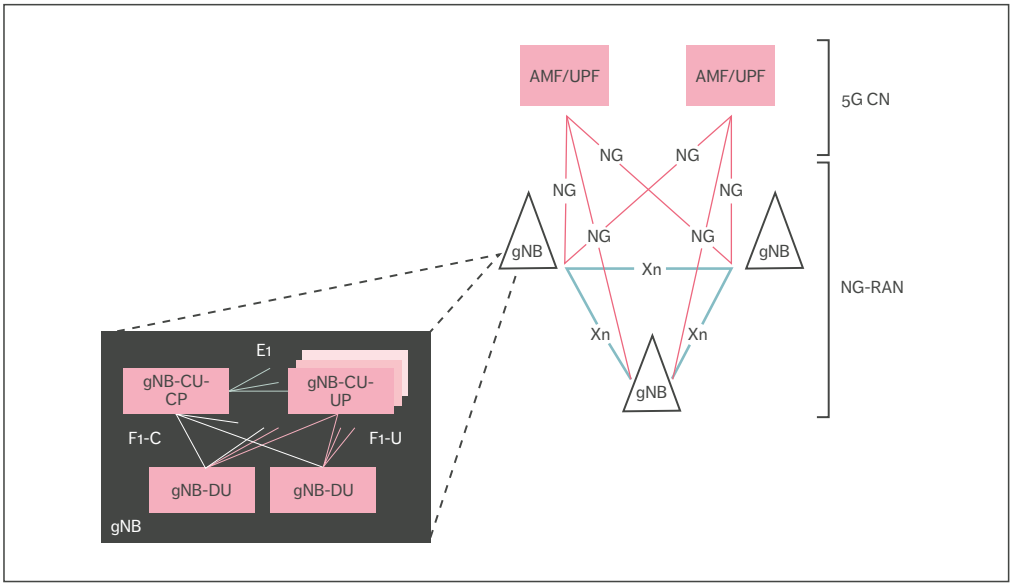


Figure 3 NG-RAN architecture

RAN architecture support

The RAN of the 5G system – known as next-generation RAN (NG-RAN) – consists of RAN nodes that serve either Evolved Universal Terrestrial Radio Access (E-UTRA) or NR cells. The bottom left corner of *Figure 3* illustrates how RAN nodes (gNBs) that serve NR cells can be split into central units (CUs) or distributed units (DUs). A DU hosts functions related to lower radio protocol layers, while a CU hosts functions related to higher radio protocol layers (RRC and Service Data Adaptation Protocol/Packet Data Convergence Protocol). Several DUs are connected to their serving CU nodes via the F1 interface, while RAN nodes may be interconnected by means of the Xn interface. A CU may be further split into a control plane (CP) part (CU-CP) and several UP parts (CU-UP).

The functional decomposition of NG-RAN nodes serves a multitude of different deployment options, including those where a CU is deployed to serve a large number of DUs corresponding to a large serving area. For example, the CU would be able to very efficiently control UE mobility while minimizing signaling traffic toward the 5G CN and between RAN nodes.

With regard to the inactive state, the CU would be able to control tasks such as the assignment of the UE's RNA based on the UE's mobility behavior and certain RAN topology knowledge. Based on Ericsson's proposal, both the CP and UP resources can remain configured in the CU when the UE is in the inactive state. The benefit of this is a reduction in processing and signaling if the UE returns to the same CU, which is highly likely in this type of deployment.

Fallback to legacy procedure

If, for any reason, the UE and the RAN end up in an unsynchronized state and the resume procedure fails, the UE will automatically switch over to the legacy idle-to-connected transition procedure that involves CN signaling. This solution could also be useful if the RAN is unable to retrieve the UE configuration or the UE configuration has been actively discarded.

> THE CN IS ABLE TO MAKE CONTACT WITH INACTIVE UEs IF THE RAN CONFIGURATION IS LOST OR DISCARDED

In these situations, the network will respond with an RRC connection setup message instead of an RRC resume message when the UE sends the RRC resume request. When the UE receives the setup message, it will discard the old RAN-related UE configuration and proceed according to the legacy idle-to-connected procedure.

UEs in the inactive state will listen for both RAN- and CN-triggered paging, so that the CN is able to make contact with inactive UEs if the RAN configuration is lost or discarded. This capability is also useful if a UE has been out of radio coverage and missed RAN paging, resulting in the RAN node releasing the UE configuration. To reduce UE battery consumption, the solution provides a mechanism for the RAN and CN to coordinate the paging occasions, so that the UE only needs to wake up once to listen to both.

Future enhancements

While the essential components for the inactive state are supported in Rel-15, there is an opportunity for further enhancements of the NR standard in later releases. There are several use cases and scenarios that would benefit from enhancements to the applicability and efficiency of the inactive state, particularly in the areas of early data transmission (EDT), early measurements and long discontinuous reception (DRX).

Early data transmission

To meet the more stringent requirements of future 5G use cases, it will soon be necessary to reduce UL latency even further. EDT is a feature that would allow opportunistic data transmission to commence during the connection resume procedure. With the resume procedure as specified in Rel-15, connection resume procedures are completed before data transmission can start. With EDT, data transfer can begin in parallel with transmission of the resume request message in the UL and the resume message in the downlink (DL). Security and radio bearers are resumed before submitting the resume request message to lower layers,which allows multiplexing of data with signaling in the Medium Access Control (MAC) layer.

EDT has already been introduced in LTE-M and Narrowband-IoT (NB-IoT), where traffic is expected to comprise the transmission of small amounts of data and one of the primary objectives is long UE battery life. For use cases where traffic consists of only one UL and/or one DL data packet, EDT improves energy efficiency by enabling the network to release the UE to the inactive state without the need for intermediate resume and resume complete messages.

Early measurements

NR Rel-15 already supports the aggregating of multiple carriers for higher data throughput using either carrier aggregation (CA) or dual connectivity (DC). When transitioning from the inactive or idle state to connected state, however, the UE only has access through one carrier. Faster setup of CA or DC would make it possible to further reduce the session setup latency. However, the usefulness of CA and DC depends on the network understanding of the radio environment.

Early measurement reporting is a feature currently being standardized in Rel-16 to improve the setup of CA and DC by enhancing NR to support early radio measurement reports during the transition from the inactive to connected state – that is, in parallel with the resume complete message. This would be possible in NR because when the UE is suspended, it receives the security parameters needed to encrypt the sensitive measurement report. When security is activated, early measurement reports can be multiplexed with a resume request or multiplexed with a resume complete message (if requested by the RAN in the resume message).

Long discontinuous reception

DRX, a feature that enables the UE to turn off its receiver, is imperative in use cases where device energy efficiency and battery life are important considerations. The longer the transmitter and receiver can be turned off, the more energy the UE can save (the longer the battery life). Long DRX has traditionally been supported in the idle state.

To enjoy the benefits of both signaling reductions and long DRX, it is desirable to extend DRX cycles in the inactive state to the same length as DRX cycles in the idle state. A key aspect of the inactive state, however, is that, from the CN point of view, the UE remains connected and DL data arriving to the CN would normally be forwarded to the RAN node serving the UE. The RAN would buffer the data until the UE is reachable.

With short DRX, the amount of data that would need to be buffered is limited. With very long DRX, however, the buffering requirements in the RAN grow and may exceed what is normally needed, which would become costly. To mitigate this and make use of (already existing) CN buffering capability/capacity, the RAN may indicate to the CN that the UE is not available for DL data while in the inactive state. In this event, the CN will buffer the data and notify the RAN, so that the RAN can inform the CN when the UE becomes available again. It is anticipated that the use of long DRX in the inactive state with buffering offloaded to the CN would further improve battery life, while enabling efficient (re)use of buffering capabilities in the network.

Conclusion

Improved connection, state and mobility handling are key requirements of many current and future 5G use cases, including smart transport and critical control of remote devices. At Ericsson, our 4G/5G research activities and lessons learned from legacy networks have enabled us to identify solutions that significantly lower both latency and battery consumption by reducing the amount of signaling required during state transitions. As a result of this work, the standalone version of the 5G NR standard includes a new Radio Resource Control state model that features a new state called inactive.

The inactive state in 5G NR is a key enabler for emerging use cases that require low latency communication and minimal battery consumption. An additional benefit of the new state is that the decreased processing effort in the network results in an increase in overall system capacity.

Rel-15 includes all the essential components for the inactive state. Future releases should focus on providing applicability and efficiency enhancements, particularly in the areas of early data transmission, early measurements and long discontinuous reception.

References

1. **Ericsson, 5G use cases, available at:** *https://www.ericsson.com/en/5g/use-cases*
2. **2016 IEEE International Conference on Communications Workshops (ICC), A novel state model for 5G Radio Access Networks, Da Silva, I.L.; Mildh, G; Säily, M; Hailu, S, abstract available at:** *https://ieeexplore.ieee.org/document/7503858*
3. **Ericsson, Handling of inactive UEs, 2016, 3GPP RAN2#94, R2-163998, available at:** *http://www.3gpp.org/ftp/TSG_RAN/WG2_RL2/TSGR2_94/Docs/R2-163998.zip*
4. **3GPP, TS 38.300, NR; Overall description; Stage-2, available at:** *https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3191*
5. **3GPP, TS 38.331, NR; Radio Resource Control (RRC); Protocol specification,available at:** *https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3197*

THE AUTHORS

Icaro Leonardo Da Silva

◆ joined Ericsson Research in 2010 and currently serves as a master researcher in radio network architecture and protocols. His work has largely focused on standardization and concept development for LTE and 5G NR, and in particular on CP topics in 3GPP RAN2, for which he was awarded the Inventor of the Year prize for 2018. Da Silva led the 5G CP in the EU project on 5G RAN architecture, METIS-II, which is part of the 5G-PPP framework. He holds an M.Sc. in electrical engineering from the Federal University of Ceará (UFC), in Fortaleza, Brazil.

Gunnar Mildh

◆ is an expert in radio network architecture in the Network Architecture and Protocols department at Ericsson Research. He joined the company in 2000 and has worked on standardization and concept development for GSM/EDGE, HSPA, LTE(-A) and 5G NR. His focus areas

include radio network architecture and protocols, and more recently 5G architecture including RAN and Packet Core. Mildh holds an M.Sc. in electrical engineering from KTH Royal Institute of Technology, Stockholm, Sweden.

Paul Schliwa-Bertling

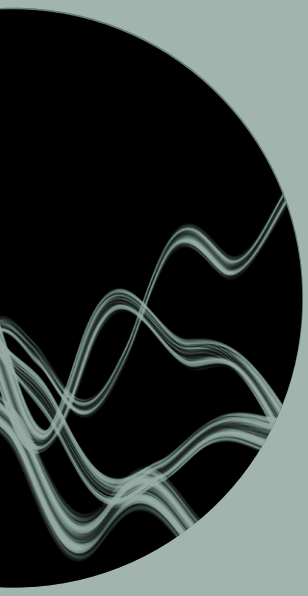
◆ joined Ericsson in 1996 and currently serves as an expert in mobile networks architecture and signaling at Ericsson Research in Linköping, Sweden. He has worked extensively with the development of RAN product and system-level concepts as well as 3GPP standardization across multiple generations of RAN and CN. His current work focuses on the evolution of network architecture and the related signaling aspects contributing to 3GPP standardization. He holds an M.Sc. in electrical engineering from the University of Duisburg-Essen in Germany.

Magnus Stattin

◆ joined Ericsson Research in 2005, where he currently serves as a principal researcher. Over the years his work has focused on research in the areas of radio resource management and radio protocols of various wireless technologies. He is also active in concept development and 3GPP standardization of LTE, NB-IoT, NR and future wireless technologies. Stattin holds a Ph.D. in radio communication systems from KTH Royal Institute of Technology in Stockholm.

Alexander Vesely

◆ joined Ericsson in 2013 after working at other major mobile network vendors for more than 20 years. He currently serves as the company's principal researcher for standardization. He has also held offices in the 3GPP for approximately eight years, and is still actively contributing. Vesely holds a Dipl.Ing. in communications engineering from the Technical University in Vienna, Austria.



ISSN 0014-0171
284 23-3330 | Uen

Ericsson
SE-164 83 Stockholm, Sweden
Phone: +46 10 719 0000