MEDIATEK

5G NR Voice Solutions Overview and Deployment Guidelines

Network Performance Considerations

Introduction

The industry has seen unprecedented growth and adoption of 5G. Compared to all its previous generations, 5G NR is the most accelerated cellular technology owing to its deployment flexibility and holistic use cases it can potentially address from mobile broadband to machine-type communications. Since 2019, the 5G deployment acceleration started with non-standalone (NSA) deployment "Option 3" for majority of carriers globally as it stands to be the quickest possible deployment in order to serve the user demands and to offer the 5G experience while established on the already mature LTE core and radio networks assistance. And as of now, more than 400 operators are investing in 5G Networks and more than 150 operators have launched 5G services commercially. Therefore, it is becoming important to address the migration from non-standalone deployment to the other 5G standalone (SA) deployment such as "Options 2" which offers a full connectivity to 5G core network, and perhaps the coexistence of various deployment in the same network.

In order to offer a wide variety of true 5G benefits and use cases, mobile network operators globally have envisioned to migrate to standalone deployment "Option 2" from non-Standalone "Option 3" of 5G. The standalone deployment would be an end-to-end solution that offers a modernization to the radio and core network connectivity. As of now around 70 operators are investing in advancing their networks to standalone deployment of 5G and the commercialization trend is expected to keep increasing going forward. While the initial deployment use cases of 5G NR focused on Enhanced Mobile Broadband (eMBB), a migration from NSA to SA now requires to study the implementation of voice solutions in order to complete the cycle of offerings to the end-user with the full potential of 5G technology.

Although data usage has exploded in the last decade but voice has remained intrinsic to any cellular technology and the same goes with 5G as well. It is to be understood that without the migration of all legacy features and services on new cellular technology the spectrum migration strategies of carriers would also be delayed which may affect the future advancements. In non-standalone deployment, VoLTE is used as a default voice solution, however in standalone deployment the equivalent of VoLTE is VoNR which is one of the possible voice solutions in 5G. Another solution is EPSFB (Evolved Packet Switched Fallback) which would require 4G layer under SA system and may be considered as interim solution keeping in mind the end goal of VoNR. At a high level, EPSFB is equivalent to CSFB (Circuit Switched Fallback) as we saw in the past when 4G was introduced on the top of 3G. The third solution is RAT Fallback (Radio Access Technology Fallback) which is mainly applicable to 3GPP defined deployment options 5 and 7, which are options not utilized in nowadays networks.

In this paper, we study several aspects of the two main 5G voice solutions: VoNR and EPSFB. The performance study in this paper is based on MediaTek's big data analysis and covers several networks deployed worldwide. Utilizing data collections of thousands of voice calls from several networks, it focuses on the performance aspects of call setup latency, breakdown of the delays, some aspects of in-call performance, and offers a comprehensive guide to network deployment considerations towards a successful migration of VoLTE to EPSFB and VoNR.

Introduction to 5G NR Voice Solutions

EPS Fallback and VoNR

Both EPSFB and VoNR are applicable to SA deployment where the User Equipment (UE) is served initially by 5G radio and core networks in addition to being registered to IMS (IP Multimedia Subsystem). However, the difference between EPSFB and VoNR comes at the radio level and the way call is handled when initiated. EPSFB call although initiates on NR but eventually falls back to LTE layer and succeeds over EPC core network that is also connected to IMS, effectively making it VoLTE call after the fallback. EPSFB call is coordinated between the 5GS (5G System) and EPS (Evolved Packet System) systems and transition of the call happens from 5G NR to the legacy LTE network after the negotiation of both UE and network capabilities, as shown in below figure 1.

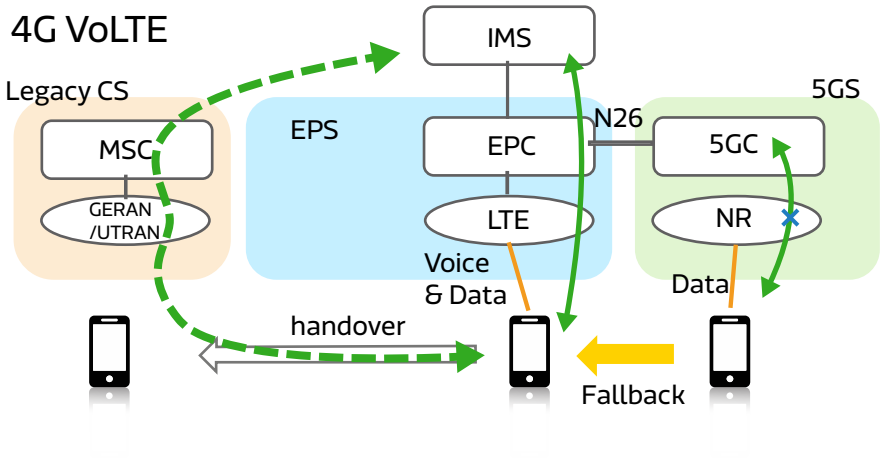
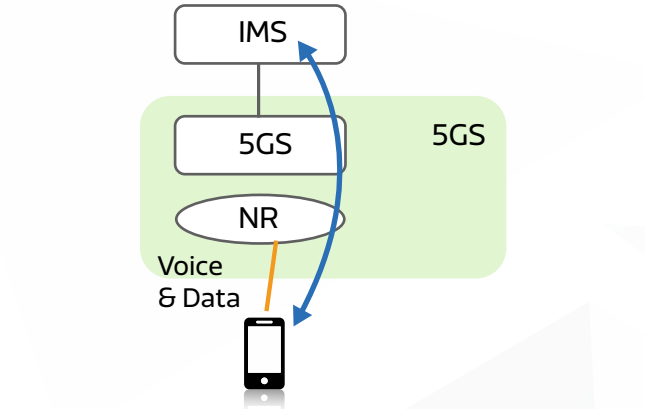


Figure 1. EPS Fallback Overview

On the other hand, VoNR in 5G is equivalent to that of VoLTE solution in LTE and the call will be handled purely by 5GS supported by IMS as shown in figure 2. The end to end VoNR network architecture remains to be the same as that of VoLTE, however it is expected that VoNR will offer better experience (considering the technology advancements in 5GC and RAN) compared to VoLTE once optimizations are in place.



EPS = EPC + E-UTRAN (E-UTRA (LTE) & NR) + UE
5GS = 5GC + NG-RAN (E-UTRA (eLTE) & NR) + UE

Figure 2. Voice over NR (VoNR) Overview

In addition, EPC may remain active as core network connectivity with UE, until 5GS deployment matches LTE EPS in cases to provide service continuity between LTE and 5G accesses or to support Voice over NR handover to VoLTE (in limited NR coverage). Therefore, the N26 interface has been introduced to be an inter-CN interface between the MME (Mobility Management Entity) in EPC and AMF (Access and Mobility Management function) in 5GC to enable interworking between them, and it is used to provide seamless session continuity, as illustrated in the end-to-end network architecture in figure 3.

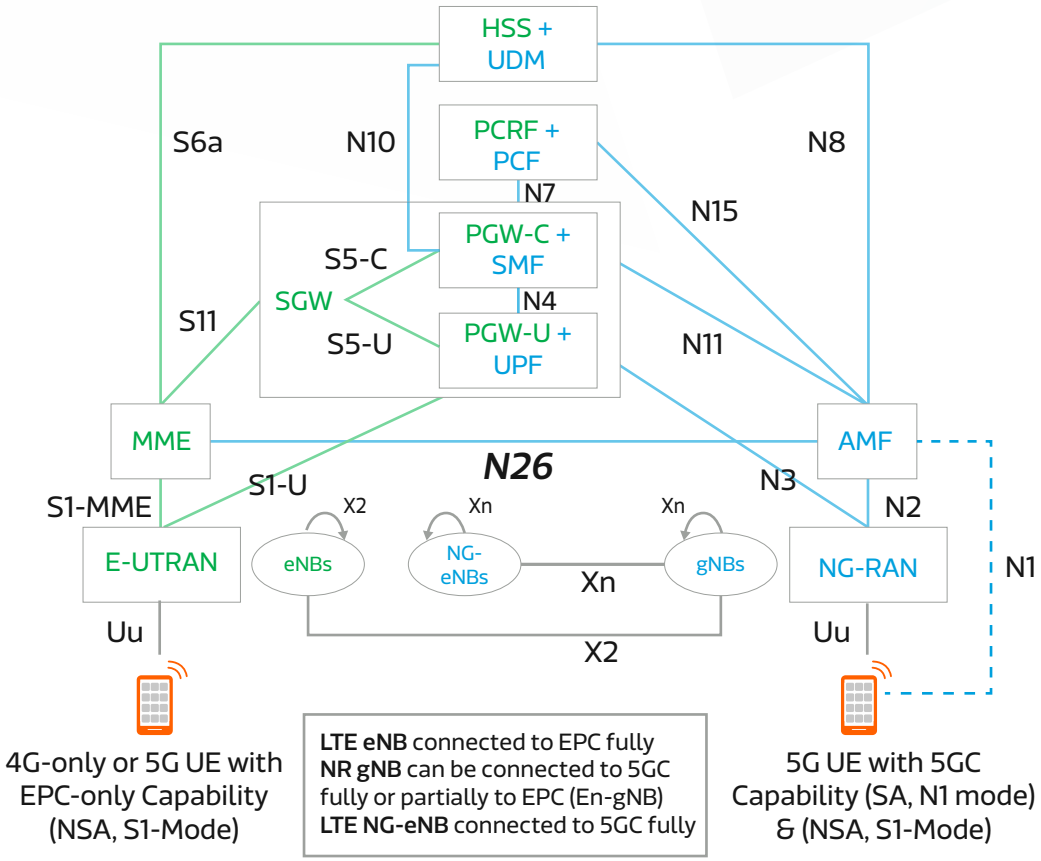


Figure 3. 5GC and EPC Interworking Architecture (Non-roaming Architecture)

The support of N26 interface in the network is optional for interworking, whereas mobility procedures such as Idle and Connected mode mobility (inter-system re-selection and handover, respectively) between EPC and 5GC are supported using N26. Therefore, for a network that supports interworking procedures with N26, the UE operates in single-registration mode only, given also that the support of single registration mode is mandatory for UEs that support both 5GC and EPC NAS. In single-registration mode, the UE has only one active MM (Mobility Management) state, and is either in 5GC NAS mode or in EPC NAS mode (when connected to 5GC or EPC, respectively). UE maintains a single coordinated registration for 5GC and EPC. For mobility in UE single-registration mode, either using or not using the network N26 interface for interworking is possible. This means that another UE registration mode exists which is dual-registration mode. In this mode, the support of N26 network interface between AMF in 5GC and MME in EPC is not required, because UE handles independent registrations for 5GC and EPC using separate RRC connections, where the UE may be registered to 5GC only, EPC only, or to both 5GC and EPC. In the case 5G network does not support VoNR/EPSFB, dual-registration mode allows the UE to register to 5GS (for data) and EPS (for voice) at same time, for which UE needs to support dual-standby radio, which can increase power consumption. When network does not support N26 interface, the delay of the inter-system change is potentially longer and ongoing voice call is affected by additional EPC attach procedure. As a summary, UE single-registration mode with N26 support by network is considered the highly preferred implementation over both single-registration without N26 or dual-registration mode owing its better performance to lower data interruption during the inter-system handover, better device power consumption with single-standby, and better voice call setup delays experience. The focus of the study in this paper is on UE single-registration mode with N26.

At the stage of UE registration, the 5GS and UE capabilities are negotiated during the initial registration process to conclude if EPSFB or VoNR can be utilized. As shown in figure 4,

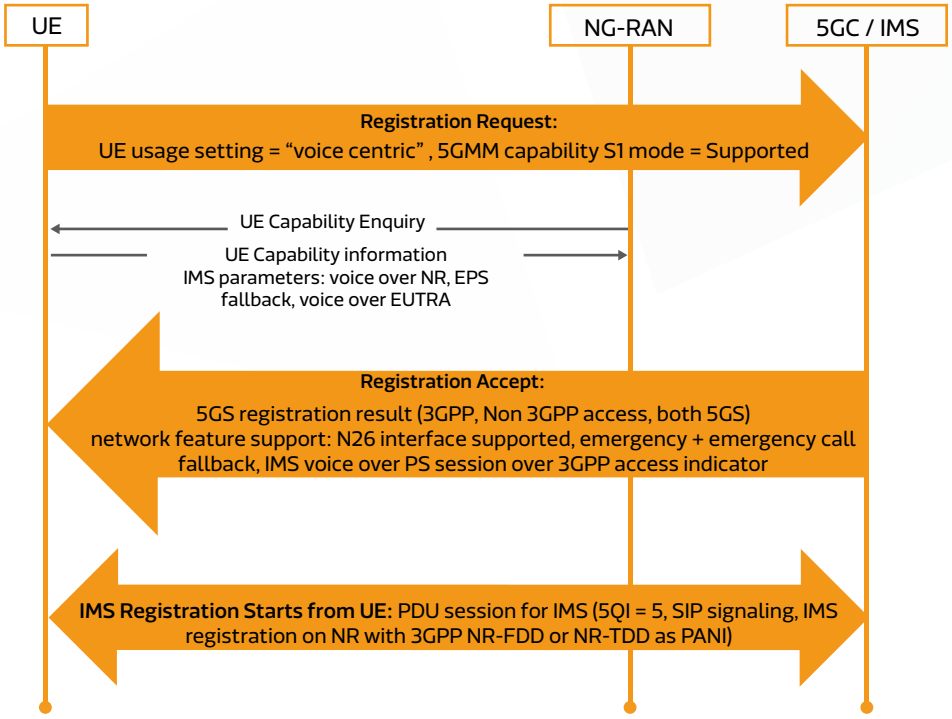


Figure 4. 5GC Registration Procedure – Voice Services

There are two level of capability negotiations:

1. The capabilities indications check is handled at NAS (Non-Access Stratum) layer, called domain selection. The UE's usage setting applies to voice capable UEs in 5GS and indicates whether the UE has preference for voice services "voice centric" over data services "data centric", or vice-versa, where:
   - Voice services include IMS voice; and Data services include any kind of user data transfer without a voice media component.
   - In case "IMS voice is not available" for "voice centric" UE this might lead to new domain selection (e.g. by disabling capabilities to access 5GS, where the UE re-selects to E-UTRAN connected to EPC first), to ensure that Voice service is possible.
   - If the UE is capable of S1 mode, there is a single UE's usage setting which applies to both 5GS and EPS .
2. To maintain the voice service in NG-RAN, the UE provides additional capabilities over RRC (Radio Resource Control) layer, that are used to determine accurate NR voice support options. In UE capability Information message in NR RRC layer, UE can convey capabilities related to IMS highlighted in table 1.

Table 1. UE capability Indicators for VoNR and EPSFB

Capability Message Parameter	Definition
voiceFallbackIndicationEPS-r16	Indicates whether the UE supports *voiceFallbackIndication* in *RRCRelease* and *MobilityFromNRCommand*. If this field is included, the UE shall support IMS voice over NR and IMS voice over E-UTRA via EPC.
voiceOverEUTRA-5GC	Indicates whether the UE supports IMS voice over E-UTRA via 5GC (i.e. RAT Fallback). It is mandated to the UE if the UE is capable of IMS voice over E-UTRA via 5GC. Otherwise, the UE does not include this field. If this field is included and the UE is capable of E-UTRA with EPC, the UE shall support IMS voice over E-UTRA via EPC.
voiceOverNR	Indicates whether the UE supports IMS voice over NR. It is mandated to the UE if the UE is capable of IMS voice over NR. Otherwise, the UE does not include this field. If this field is included and the UE is capable of E-UTRA with EPC, the UE shall support IMS voice over E-UTRA via EPC.
voiceOverSCG-BearerEUTRA-5GC	Indicates whether the UE supports IMS voice over SCG bearer of NE-DC

After voice over IMS is determined, UE can start IMS registration and then be able to make VoNR (or EPSFB) calls. At the call initiation, the call setup flow differs between VoNR and EPSFB as illustrated in figure 5.

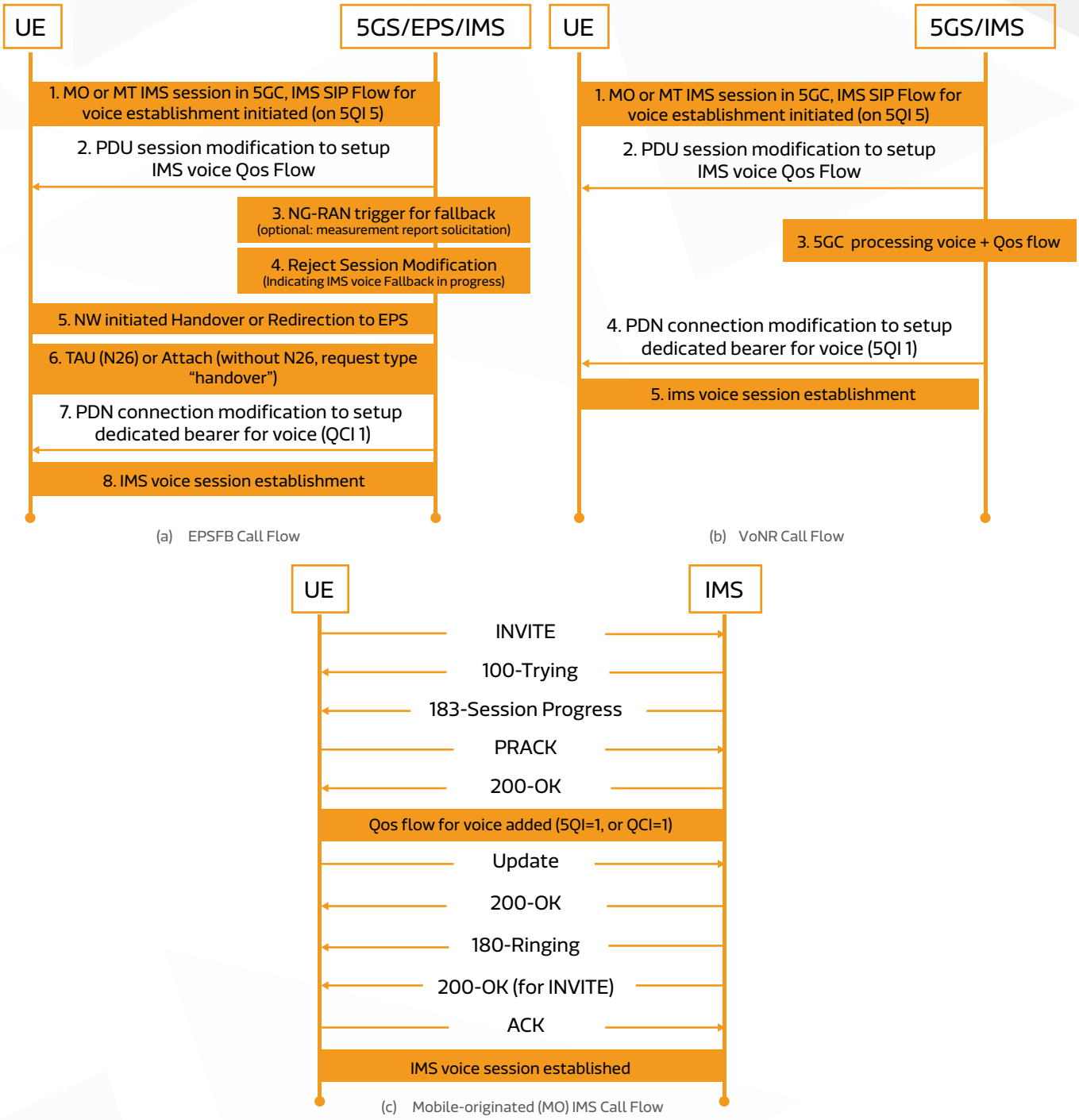


Figure 5. End-to-End Call Flow for EPSFB and VoNR

In figure 5(a), if a request for establishing the QoS flow for IMS voice reaches the NG-RAN, NG-RAN is configured to support EPS fallback for IMS voice and decides to trigger fallback to EPS, taking into account UE capabilities, indication from AMF that "Redirection for EPS fallback for voice is possible", network configuration (e.g. N26 availability configuration) and radio conditions, then the a redirection or handover procedure to LTE starts. After the UE camps successfully on LTE cell and initiates Tracking Area Update procedure (TAU) or a fresh attach process (in case of TAU failure or no support for N26), the call continues normally as VoLTE call. In contrast, the VoNR call flows in figure 5(b) shows that the IMS call continues all the way on 5G system without any inter-RAT system change interruption. Therefore, the network is the one controlling the initiation of EPSFB after knowing the UE capabilities for voice services over 5G system. Finally, figure 5(c) shows an indicative IMS SIP (Session Initiation Protocol) call flow between the UE and IMS server during Mobile-originated call, with SIP messages used in later sections to calculate the call setup latency.

In the next sections, we discuss the deployment observations and challenges related to VoNR and EPSFB, and some of the possible solutions to overcome those challenges.

VoNR Performance Analysis

Data shown in table 2 represents the overall MO (Mobile Originated) IMS call setup latency observed on various networks involving VoNR, EPSFB and VoLTE. Note that values and call setup latency can vary from one network to another depending on actual network design (band, bandwidth, coverage, etc...). Therefore, the purpose of showing the data is not to benchmark the actual values, but to take the relative differences in order to establish a good analysis model of the technologies under evaluation. For the sake of the performance analysis in this paper, 18000 voice calls were processed, but it is to be noted that the sample size for VoNR in particular is less than EPSFB and VoLTE because of the limited commercial deployment availability at the time the paper is published.

The MO call setup latency is calculated from SIP_Invite to SIP_180_Ringing, which are SIP messages shown in figure 5(c). The analysis is focused on MO calls because it is inclusive of MT call delays as well. The statistics cover different test cases and the analysis is run over all scenarios (mobility and stationary conditions).

Table 2. MO call setup time in different voice solutions

MO Call Scenario	Call Setup Latency from Connected Mode (sec)			Call Setup Latency from Idle Mode (sec)		
	Mean	Median	Mode	Mean	Median	Mode
VoNR	4.02	3.20	2.60	4.64	3.80	3.40
EPSFB	4.96	4.65	4.00	5.17	5.00	4.80
VoLTE	3.55	2.60	1.60	4.24	3.00	2.60

It is observed that VoLTE call setup is generally optimized across major networks due to the legacy implementations while VoNR and EPSFB may require additional effort to come in line with VoLTE in an end-to-end-optimization. Next, we will analyze the data to break down the areas where VoNR improvement may be needed, and where in particular EPSFB is showing higher call setup latency than VoNR and VoLTE.

Starting first with the VoNR performance analysis, the MO call setup latency can be broken down further into Delay_1 and Delay_2 which are defined as follows:

- Delay_1 is referred to as the Call Access Delay (from MO SIP_Invite till SIP_183_Session_Progress) and defines the IMS processing delays between MO and MT for e.g. MT being paged, MT setting up IMS call or say MT side core delay for setting up dedicated bearer.
- Delay_2 is referred to as the Call Processing Delay (from SIP_183_Session_Progress till SIP_180_Ringing) and defines the radio delays also where in this period, NR in VoNR or LTE in VoLTE air interface messages flow between MO UE and Network in addition to EPS/5GC core network delays for MO side Dedicated Bearer Setup till the final step of establishing the call between MO and MT in the ringing stage.

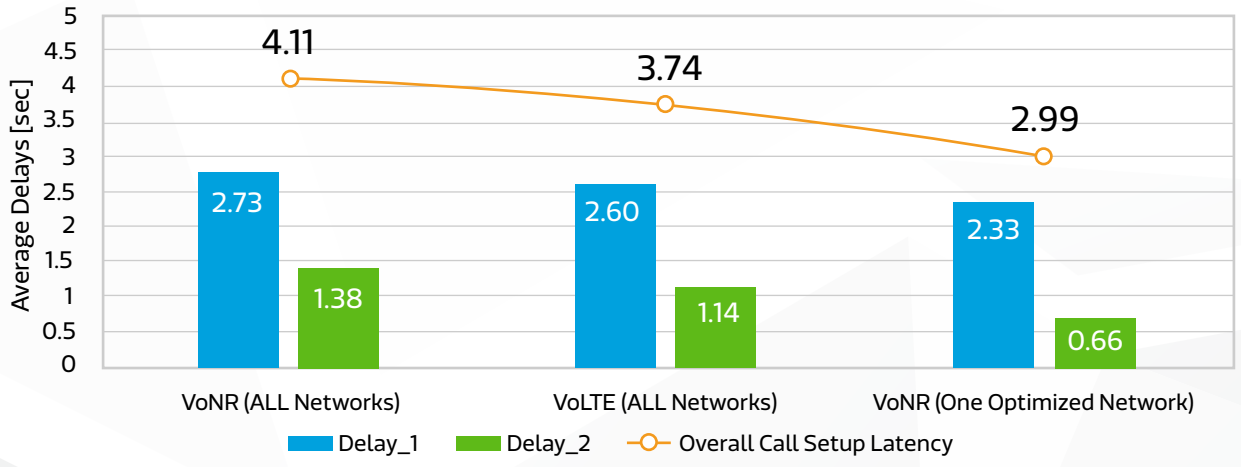


Figure 6. VoLTE and VoNR Call Setup Latency

As depicted in figure 6, it can be observed that the major delay comes from Delay_1, and can be further concluded that there is a scope of improvement in radio and core if we compare the data with that of VoNR Optimized Network. In addition, if we compare the average values of Delay_2 of all networks to that of one VoNR optimized network, it can be seen that the average delays are more than double. At a high level it is safe to assume that perhaps the VoNR deployment will also observe several phases of optimization as it happened during VoLTE deployment phase. It is not unrealistic to say that with more optimizations the VoNR performance will be matched to that of VoLTE, and in future VoNR may even outperform due to the enhancements that come with 5G core and radio. Generally, to illustrate this potential improvements of VoNR radio and core network processing, we further looked into IMS signaling delay analysis. We defined "Abs. IMS SIP Latency" as IMS SIP Round Trip Time, calculated as the average time difference between SIP messages (e.g. between SIP INVITE on UL ↔100 Trying on downlink, Update ↔OK, etc..). It is a procedural delay between UE and IMS server that covers the IMS SIP path over the radio/core network all the way into IMS server (E2E latencies on 5QI=5). As illustrated in figure 7, due to NR enhancements in terms of radio and core delays (e.g. link quality and advanced antenna techniques), VoNR IMS Signaling delay reduced by ~42% compared to VoLTE, while the increased EPSFB IMS latency (of ~9% compared to VoLTE) is due to the radio/core switching latencies during the process. This indicates that potentially VoNR IMS signaling performance is better than VoLTE and EPSFB and it can possibly reduce the IMS failures due to timeout procedures, improve RTP timeouts and the overall Jitter, and hence having better coverage (delay budget) where the call quality could also improve.

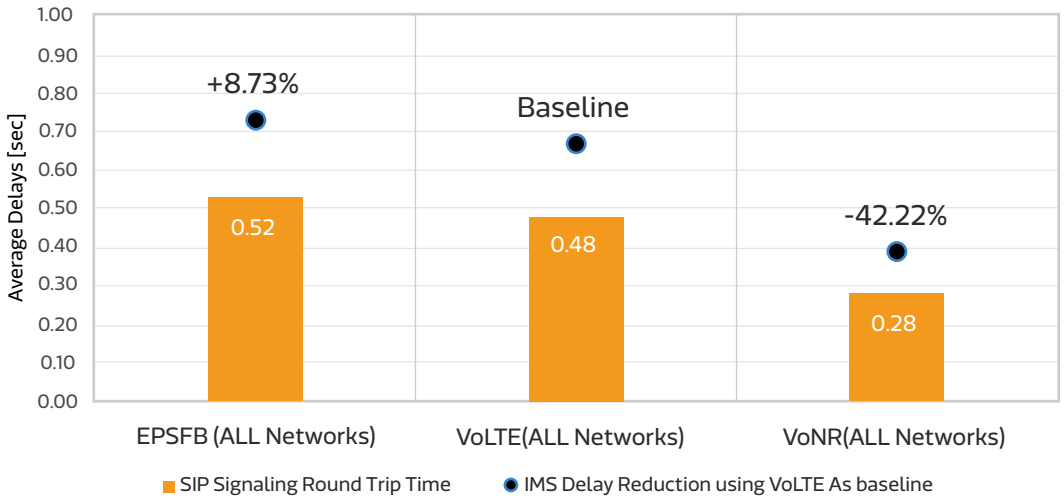


Figure 7. VoNR vs. VoLTE vs. EPSFB: IMS Signaling Delay Analysis

Now, coming back to the areas of optimization for the VoNR network shown in figure 6 (on the left hand side), if we further breakdown the Delay_1, one of the areas that require attention is "Paging Delays". MO UE will only continue with the call setup as long as MT UE is paged successfully, and hence Delay_1 is a reflection of how quickly MT UE can be paged. The default paging cycle can definitely affect the Delay_1 values, figure 8 below shows the effect of paging cycle on call setup time.

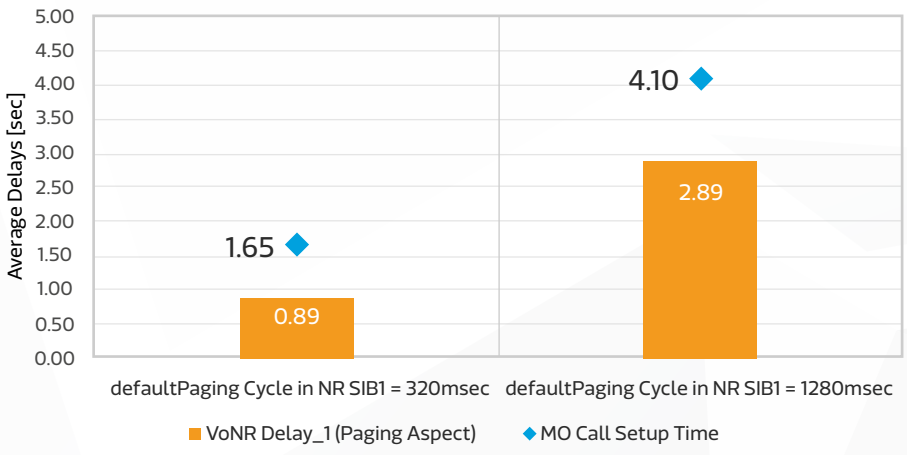


Figure 8. Impact of paging cycle (Idle mode DRX cycle) on Call Setup Latency

As observed, with paging cycle of 320msec, the average time for MO UE to establish the IMS call in between SIP INVITE to Session Progress (this is the paging period at the MT UE, and it is the perceived delay at the MO side) is ~2 sec lower than the case where the paging cycle is configured in NR SIB1 as 1280msec mainly because of the UE's shorter sleep cycle. While the paging delays will be better with shorter default paging cycle but the tradeoff is always between power consumption vs. paging delays, thus it becomes essential to improve the paging success rate even when the paging cycle is configured with higher value. Let us briefly understand the paging mechanism and UE behavior in LTE and NR which will be helpful in understanding the possible solutions.

In LTE when the UE wakes up after the sleep cycle of 1.28 seconds (which is the typical DRX cycle in LTE), the UE monitors the cell reference signal, CRS, which is always on and transmitted every subframe. LTE CRS is transmitted in every subframe and hence Paging Occasion (PO) in the Paging Frame (PF) can be decoded immediately after UE synchronizes to the serving cell and measures best neighbor cells for re-selection, minimizing wakeup time in every DRX cycle. The UE typically wakes up 3ms before the paging occasion in order to estimate the channel radio condition of the serving cell and Intra frequency neighboring cells. After the UE reads the paging occasion and in case there is no subsequent paging message, UE may continue with Inter-frequency measurements in case serving cell is below Snonintrasearch followed by sleep cycle as shown in figure 9.

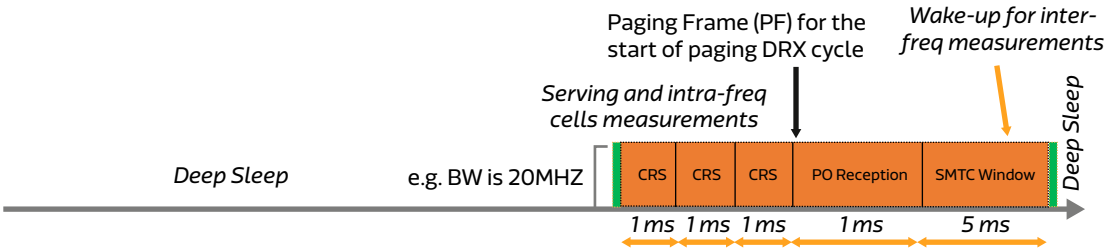


Figure 9. Paging Procedure and Idle Mode Measurements in LTE

However in NR, the mechanism is different than that of LTE as there is no per-slot reference signal structure and instead synchronization signal referred to as SS/PBCH Block (SSB) is used for channel measurements. It should be noted that SSB is typically transmitted in the period of 20 msec. Unlike in LTE where UE wakes up 3 msec before the paging subframe, UE in NR could potentially need to wake up 60 msec before the paging occasion in order to identify the SSB of the serving cell and neighboring cell required to perform measurements, as explained in figure 10. Also the paging occasion in NR can depend on the SSB beam concept. As a result, more power will be spent on paging monitoring because of extended duration of occasion and also because of the time spent before paging occasion. This difference in paging structure itself costs ~70% more power in NR than in LTE when it comes to paging procedure of UE[1].

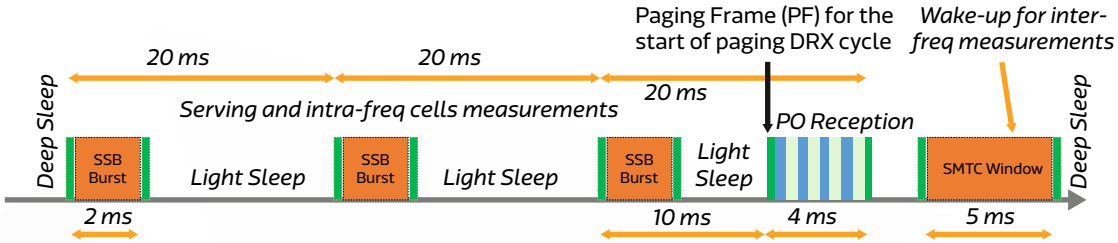


Figure 10. Paging Procedure and Idle Mode Measurements in NR

Therefore, reducing the default paging cycle to shorter value such as 320 msec from 1280 msec in NR may affect the UE adversely and will deteriorate the power consumption. Although shorter DRX cycle can improve the latency as shown in figure 8 but it is not recommended because of the factors discussed above. Some of the paging enhancement solutions are studied in Rel. 17 that offer adequate balance between paging delay and UE power consumption and makes lower default paging cycle possible (for e.g. 320 msec). These techniques proposed in 3GPP Rel. 17 can potentially improve power consumption by 25%, as follows:

- Paging Early Indication (PEI): Network sends an indication to UE before a Paging Occasion (PO) and the UE is not required to decode and process the information if the negative indication is received. The UE can skip not only paging PDCCH/PDSCH monitoring but also avoid SSB processing unless cell reselection is required which will significantly reduce the power consumption.

[1] Evaluations and Observations for R17 UE Power Saving Enhancements in 3GPP R1-2003667

- UE Grouping: The idea is to divide UEs monitoring the same PO into subgroups in order to reduce the false alarm rates, in other words the objective is to increase the probability of UE decoding the message and identifying that the decoded message contains page for this UE. This helps in minimizing occasions for UE to read/decode the page.

Another area of evaluation is the RRC state transition during the call setup and how it adds to overall delay. It was observed that the average VoNR MO call access delay is ~450 to 700 msec higher from Idle compared to the calls accessed from connected mode as shown in figure 11. This access delay for Idle mode typically includes RRC connection establishment + network sending 100_Trying + setting up other procedures like RRC Reconfiguration for measurements, Security (radio/core), dedicated bearer (core), and waiting time for MT to respond to the call. Whereas, the connected mode UE experiences only the core and MT response delays as the connection has been established already beforehand.

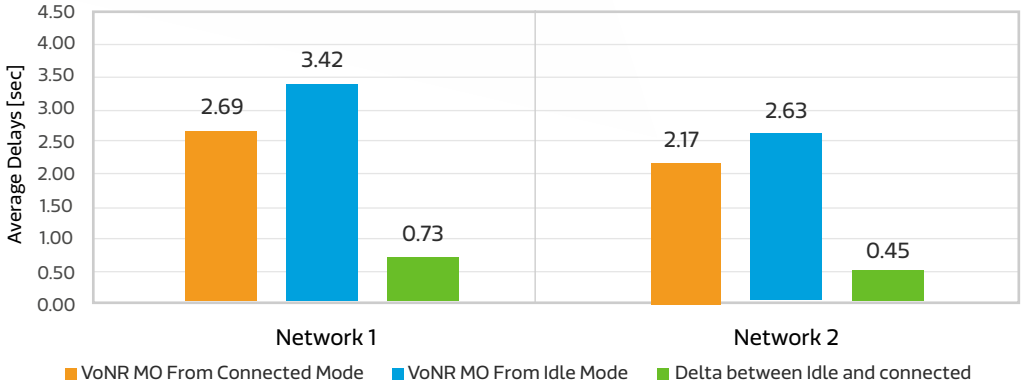


Figure 11. Delay_1 latencies from different RRC states (From SIP_INVITE till SIP_Session_Progress)

Therefore, the newly introduced RRC state "RRC Inactive" state in 5G NR can potentially bring delay improvement in this situation while transitioning to connected state during the voice call setup. During the transition from RRC Inactive to RRC connected, the UE may not have to follow legacy procedures as it used to happen from RRC idle to RRC connected. Instead, the transition from RRC Inactive can simply just resume the connection without having to establish a new RRC Connection. This way, the additional RRC messages exchange could be avoided which eventually will boost the call setup time, in our study we observe that such gains could be close to 58% or more as shown in figure 12. The latencies shown are described as follows:

- RRC Setup Latency: From RRC Connection Request or RRC Resume Request to RRC Connection Complete or RRC Resume Complete. From Idle, this delay involves only RRC Connection establishment between UE and gNB. From Inactive, the target cell resumes the stored configuration and applies any necessary modifications such as the configuration of measurements or UE Context Retrieval procedure (depending on the mobility scenario).
- Connection Setup Latency: RRC Connection Request or RRC Resume (fallback case) till RRC Reconfiguration Complete (i.e. till the point at radio side that completes the procedure RRC establishment was intended for). From Idle, it involves signaling between UE and gNB including Core network such as Service Request, Security setup (UE context setup for security and bearers). From Inactive, in most cases, UE moves directly to connected mode without additional signaling. RRC resume also has the possibility of using delta signaling, in which only changed parameters are signaled. This option is not possible for UEs in the idle state. Without additional signaling, even more than 58% (up to 75%) delay reduction can be achieved compared to the transition from idle mode.

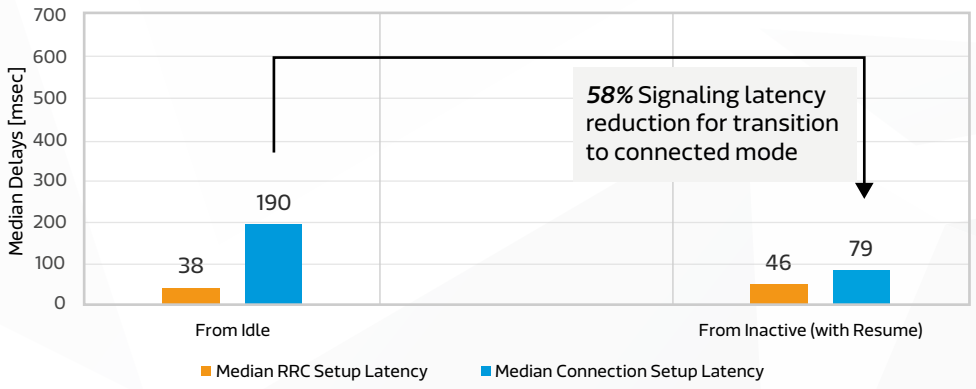


Figure 12.Access Signaling Latency to RRC Connection Mode

EPS FB Performance Analysis

EPSFB is another type of voice calls that are initiated from 5G in the case VoNR is not possible. As shown in the call flow in figure 5(a), one of the essential aspects of EPSFB is the way the NR to LTE inter-RAT transition is triggered (at step 5 in the figure). Generally, there are three types of EPSFB implementations across the network:

- **EPS FB with Handover**: gNB configures UE with event B1 or B2 to measure LTE cells, followed by *MobilityFromNRCommand* RRC message for the NR to LTE IRAT handover, immediately after UE sends the measurement report.
- **EPS FB with Redirection and without LTE Measurement**: gNB sends *NR_RRCRelease* with *redirectedCarrierInfo* to UE for the NR to LTE IRAT, immediately after the IMS call is processed. Typically used when the LTE coverage is always available.
- **EPS FB with Redirection and with LTE Measurement**: gNB configures UE with event B1 or B2 to measure LTE cells, followed by *NR_RRCRelease* with *redirectedCarrierInfo* to UE for the NR to LTE inter-RAT, immediately after UE sends the measurement report. It is typically used in cases there can be multiple LTE carriers deployed and network prefers for UE to measure the cells to direct UE to the best frequency layer.

EPSFB with Handover is the most common implementation observed during the statistical analysis we run (66% utilization) followed by Redirection without LTE Measurement (31% utilization) and Redirection with LTE Measurement 3% utilization) over all the MO calls analyzed in this paper. Figure 13 explains the EPSFB delays breakdown using two categories to illustrate the latencies:

- Delay_1 (Measurement Delay): is the time in NR, calculated from UE sending SIP_Invite till the UE receives RRC message for Handover or Redirection to LTE.
- Delay_2 (LTE Signaling Delay): is the time UE spends in LTE, calculated from last handover or or Redirection RRC message till SIP_180_Ringing.

It is observed that EPSFB using the technique of “Redirection without LTE Measurement” has the lowest overall call setup latency, an average reduction of ~7%, comparing with “Handover EPSFB” and “Redirection with LTE Measurement”.

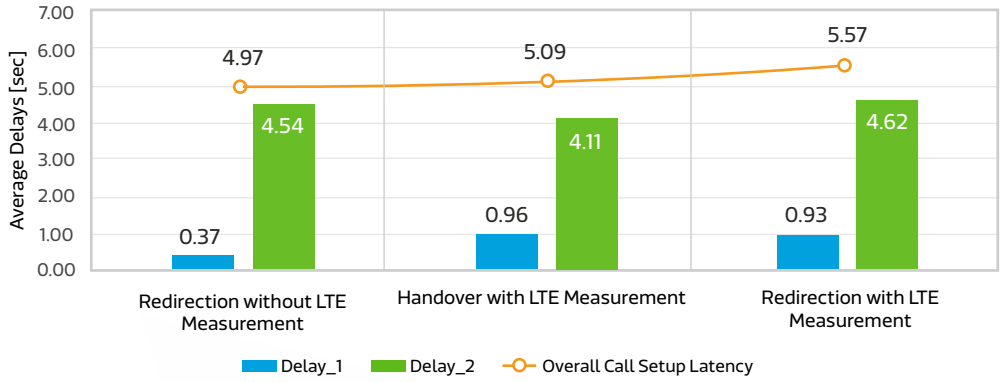


Figure 13. EPSFB Call Setup Latency Breakdown with Different Techniques

In the case of “Redirection without LTE Measurement” the UE does not need to perform any LTE measurements, and the call setup latency is mainly due to UE waiting for RRC Release during IMS call processing in 5GC. Avoiding LTE measurement can give UE a shorter call setup time, but at the same time may lead to lower call setup success rate, depending on the coverage planning overlay between NR and LTE. Hence a balance needs to strike between the tradeoff of shorter call setup time and lower call setup success rate.

The case of “Redirection with LTE Measurement” has the longest delay as new RRC connection needs to be established when UE being redirected from NR to LTE. UE has to first move into idle mode in LTE, then re-establish a RRC connection followed by a normal VoLTE call flow establishment. Additional delay comes from the measurements needed in 5G prior to the redirection. Finally, the UE performing EPSFB with Handover does not need to go into idle mode after the transition to LTE, Instead, UE can directly be moved into LTE connected mode.

Delay_1 can be mainly observed in Handover and “Redirection with LTE measurement” implementations. From figure 14, the measurement delay in NR can be further breakdown into:

- SIP_Invite till Network Configures Inter-RAT Measurements (event B1).
- Inter-RAT measurement and reporting in NR.
- Inter-RAT Execution in NR.

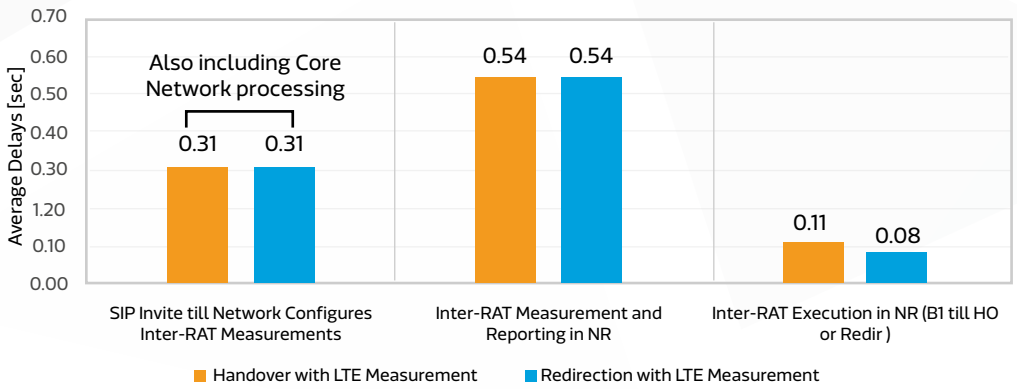


Figure 14. Breakdown of Measurement Delays in NR

In both implementation, gNB configures UE with B1 measurements (in RRC Reconfiguration) in average after 310 msec from when UE sends SIP Invite. This is the period 5GC performs PDU Session Modification rejection for EPS FB. Note that from figure 13, in "Redirection without measurement", Delay_1 = 370ms which includes both this 310ms delay and ~60ms redirection execution from gNB side. The execution time for NR to LTE Handover and Redirection is 110 msec and 80 msec. Meanwhile, the UE takes ~540 msec to measure LTE cells and then sends B1 measurement report after Time-to-Trigger (TTT) and event thresholds for B1 are all satisfied (note: all the handover and redirection measurements are configured by gNB based on event B1 where Inter RAT neighbor becomes better than a configured threshold by the network). The majority value of TTT observed in EPSFB is 320 msec. It means that the UE has to wait 320 msec before it can trigger the Handover or Redirection. Therefore, reducing TTT can shorten the delays bring by Inter-RAT measurement and reporting in NR. Reducing TTT value from 320 msec to 40 msec (a typical value used in VoLTE for Inter-RAT measurements such as eSRVCC), the EPSFB measurement delay can be reduced in one-side (MO or MT) from 540 msec to 260 msec, in average, improving the call setup latency for EPSFB, and leads to (540-260)*2 = 560 msec call setup latency saving in two-side (in the case network configures both MO and MT with 40 msec EPSFB B1 TTT). This effect was observed clearly in the test when comparing networks using 320 msec and 40 msec event B1 TTT, and in average, the EPSFB call setup latency was recorded as 5.12 msec and 4.44 msec respectively.

On the other hand, Delay_2, can be further categorized into two stages:

- The time difference between when UE Leaves NR till it camps successfully to LTE, and
- LTE camping success till UE receives SIP_180_Ringing.

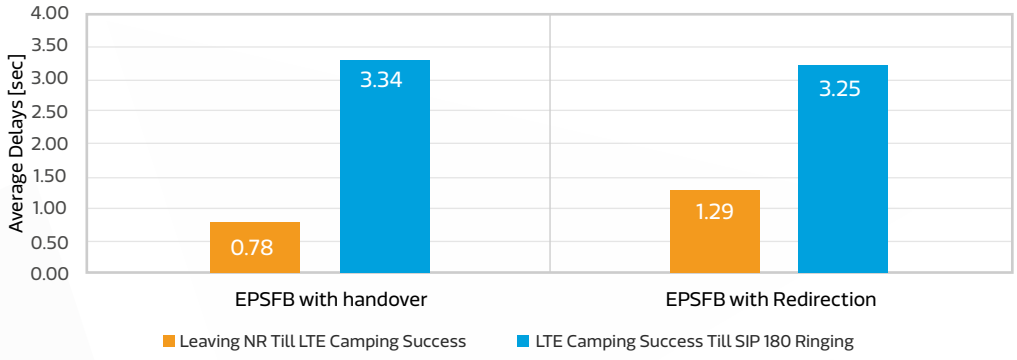


Figure 15. Breakdown of LTE Signaling Delay

The delay from when UE leaves NR till it camps successfully to LTE depends on the type of EPSFB. For EPSFB with Handover, the 0.78 sec represents mainly LTE cell acquisition including MIB/SIB + LTE Tracking Area Update (TAU) procedure (incl. EPS bearer setup), which is shown in figure 5(a) at steps 5-6. For EPSFB with Redirection, the 1.29 sec represents mainly LTE cell acquisition including MIB/SIB + LTE RRC Connection Setup + LTE Tracking Area Update procedure (incl. EPS bearer setup). Therefore it is higher for redirection for the reason of extra search (unknown cell as the redirection takes place at frequency level) in addition to setting up LTE RRC Connection, as in Handover, the UE moves directly to RRC Connected Mode in LTE.

Besides that, evaluating the LTE Tracking Area Update (TAU) procedure delays and its effect to EPSFB shows that most of the time for EPSFB with Redirection, the network adds Authentication and Security Mode procedures which adds to the delay during the NR to LTE transition, while this delay most of the time is not found EPS FB with Handover. This is typically an area related to network implementation and can provide good improvements once reviewed how frequent these procedures needed and how different it is in case of redirection and handover. In the case TAU triggered without additional Authentication and Security Mode procedures, the procedural delay affects the call setup by an additional 280 msec, while in the case of TAU with Authentication and Security Mode procedures, the additional delay goes all the way up to 670 msec.

Figure 15 also shows that the delay from when UE camps successfully to LTE (after TAU is completed) till SIP_180_Ringing is 3.34 sec for Handover and 3.25 sec for Redirection. This delay is very similar between Handover and Redirection, because it is not really affected by the radio overheads, rather it is related to EPS and IMS signaling. The overall delay after TAU procedure includes the events of EPS Bearer Context setup (QCI-1) and IMS SIP message exchange in LTE between 183_Session_Progress till 180_Ringing. This delay after TAU procedure may essentially be similar to the overall VoLTE call setup time shown in figure 6, where the IMS and core network procedures exchange takes place in LTE at this stage.

Up to this point, we have analyzed the EPSFB call setup latency and the major contributor to the delays during a successful call setup. However, there are cases where the call setup increases significantly reaching abnormal average observed as 9.51 sec, which is higher than the average observed in figure. 13. When analyzing the abnormal call setup, it was observed it is mostly coming in the occasions during EPSFB handover with measurements, where UE may be unable to find LTE cell or the handover execution fails at the gNB side. Another occasion observed is where the redirection takes place to LTE, and the UE cannot find suitable LTE cell on the redirected band. These measurement failure cases are common especially in early deployment where the coverage areas of 5G and LTE are different especially that both 5G and LTE are deployed in different bands. There are several solutions to this abnormal call setup, such as:

- Network radio planning optimization: Inter-RAT neighbor cell planning especially in cases LTE is deployed on several bands.
- Network algorithms: in case of handover failures within a certain period (e.g. 3 sec) where UE does not return any Inter-RAT event B1 measurements in 5G for LTE cells, the gNB can trigger redirection instead. This method was also utilized in CSFB.
- 3GPP Rel-16 enhancements to EPSFB: In Rel-15, there was no indication that the handover or the redirection from NR is related to EPSFB. Therefore, there is no way for E-UTRAN to prioritize the UE in the voice fallback from NR. In addition, failure handling of EPSFB may trigger UE to revert back to NR, e.g. when [NR to LTE] handover fails in LTE, UE may revert back to NR, causing EPSFB call to fail, or additional delays for UE to push the call back to LTE.

3GPP Rel-16 introduced new flag *voiceFallbackIndication-r16* which can be sent by gNB in both NR *RRCRelease* with redirection and *MobilityFromNRCommand*. This flag gives the UE the ability to handle handover failure from NR to LTE, or prioritize EPSFB call in LTE RRC Connection, as explained in figure 16. This change is related to NR side during the redirection/handover, where if the UE does not succeed in establishing the connection to the target radio access technology (LTE), and *voiceFallbackIndication* is included in *MobilityFromNRCommand* message, then the UE may take different implementations based on the following cases:

- Case 1: Suitable cell found in LTE after the handover failure
  - For example: handover to a cell found in NR but no more valid after UE moves to LTE – possibly due to delayed handover execution in NR to an LTE cell after UE reports multiple ones
  - UE performs the actions upon going to RRC_IDLE with release cause 'RRC connection failure', which would allow UE to initiate RRC Connection Procedure in LTE instead of reverting back to NR
  - This helps to sustain EPSFB call setup instead of possibly observe EPSFB call setup failure (even though with higher setup delays). It also adds better way for the Network to monitor EPSFB KPIs at gNB/eNB level
- Case 2: Suitable cell not found in LTE after the handover failure (e.g. bad LTE radio conditions)
  - UE reverts back to the configuration used in the source Pcell (NR cell) and the UE then initiates RRC connection re-establishment procedure
  - It is up to gNB to decide how to handle such handover failure, but it would probably trigger another LTE cell measurements (e.g. to another LTE carrier) or declare EPSFB failure. But has to be done fast enough by avoiding paging timeout at MT side that can an eventual failure

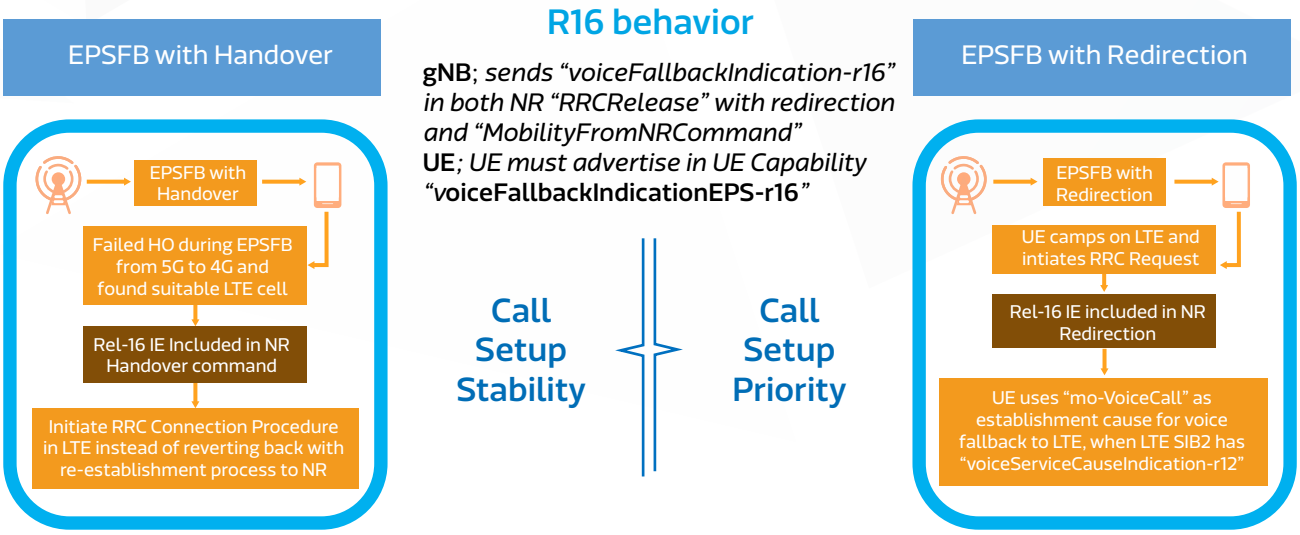


Figure 16. Enhancements to EPSFB in 3GPP Rel-16

On the other hand, as shown in figure 16, LTE Rel-16 also introduced new procedures if the UE supports *mo-VoiceCall* establishment cause and EPS fallback for IMS voice was triggered in NR via *RRCRelease* with *voiceFallbackIndication*

- The UE shall use *mo-VoiceCall* as establishment cause, in case of RRC Connection Request after redirection from NR for the purpose of voice fallback to LTE, and *SystemInformationBlockType2* includes *voiceService-CauseIndication* (LTE Rel-12 IE for IMS) and the establishment cause received from upper layers is not set to *highPriorityAccess* or *emergency*
- It also helps eNB to prioritize the return to NR after EPSFB call is over, which speeds up UE camping back to 5G

Next, we can assess the time delay for the UE to return successfully from LTE to 5G after the EPSFB is completed. This case can be of interest especially from end-user experience in order for the user to see 5G icon as quick as possible after the EPSFB call is terminated. There are several types of UE returning back to NR when call ends in LTE, and here we focus on the network side trigger for this return:

- eNB releases RRC with redirection back to 5G (direct return to NR)
- eNB performs handover from EUTRAN to NR based on measurements (note this still requires UE to do NR registration after going back to NR)
- eNB releases RRC without redirection (normal RRC Release after UE inactivity timer expires in gNB) which means UE needs to do re-selection back to NR from LTE idle mode, or trigger autonomous UE-based return to NR (as explained above)

eNB releases RRC with redirection back to 5G is the most commonly used method by networks during the analysis done where this method is utilized 91% of the returns to 5G. Handover based return to NR was not observed in any test done. Return to 5G after EPSFB latency is calculated from the point EPSFB Call released (e.g. SIP Bye) to the receiving of Registration-ACCEPT in NR The overall LTE-to-5G-return delay observed is 2.35 sec in average, where 1.5 sec is spent in the duration from when an LTE redirection is triggered though *RRCRelease* till NR *RegistrationAccept*, and the remaining time was the time eNB needed to execute the redirection after the call was terminated (from SIP BYE till *RRCRelease*). This delay shows that the user experience in seeing 5G icon is unaffected after the EPSFB call is completed.

Other VoNR Features Considerations

The paper so far has addressed the details of call setup latency. However, during VoNR call, there can be other aspects of improvements that can be needed to improve user experience such as:

- Maintain Quality of Experience for voice services. In this area, the network operators may need to ensure the user experience is at the same level with VoLTE services when it comes to call drop rate, supplementary services, call hold/swap/merge, IMS server error handling (4xx/5xx errors), call continuity in cell-edge, inter-operator and roaming calling. This area is not covered in this paper, but brought up to be an essential area of testing before the launch of VoNR services.
- Maintain Quality of Service. The end-user experience is subject to factors such as battery consumption during VoNR calls, call quality in different radio coverage (indoor and outdoor). In addition, the network capacity aspects need to consider user dimensioning for VoNR similar to what was done in VoLTE.

For battery consumption, 3GPP in Rel-16 has addressed this area with several new features that can be summarized in table 3. Some of them can have direct influence to VoNR performance especially BWP adaptation[2].

Table 3. Brief Overview of 3GPP Rel-16 Device Power Saving Framework

3GPP Rel-16 Feature	Brief Description
Enhanced cross-slot scheduling	In Rel-15, A-CSI-RS slot offset (K0) is 0 for Sub-6, and RF and RS buffering are still needed. For this, Rel-16 allows slot offset (k0) > 0. Rel-15 uses BWP switching by RRC configuration to adapt same/cross slot scheduling, while Rel-16 specifies a new adaptation within an active BWP. It saves power consumption when the data arrival is sparse.
Wake-up Signal (WUS)	A wake-up signal indication (WUS) conveyed by a new PDCCH format to inform UE whether or not to start the DRX-onDuration timer for the next DRX cycle. Beneficial in the case of sporadic traffic
Adaptive MIMO layer in BWP framework	In Rel-15, the DL maximum number of MIMO layers is configured per serving cell which is common to all the DL BWPs of the carrier. In Rel-16, the DL maximum number of MIMO layers can be separately configured for each DL BWP. The DL maximum number of MIMO layers can be changed through BWP switch which can reduce power consumption by adaptation to less number of receive antennas at UE.
Enhanced UE-assisted information (UAI)	Challenging for network to customize configurations for all the devices based on their power saving and overheating protection requirements. Supported UE-assisted information (UAI) in Rel-16 are as follows: • If UE prefers an adjustment in the connected mode DRX cycle length, for the purpose of delay budget reporting; • If it is experiencing internal overheating; • If it prefers certain DRX parameter values, and/or a reduced maximum number of secondary component carriers, and/or a reduced maximum aggregated bandwidth and/or a reduced maximum number of MIMO layers and/or minimum scheduling offsets K0 and K2 for power saving purpose; • If it expects not to send or receive more data in near future, it can provide its preferred RRC state.
Second DRX configuration	In Rel-15, in FR1+FR2 CA, all the cells share the same single C-DRX configuration and timing. In Rel-16, secondary C-DRX can be applied to second group of cells (e.g. FR2 cells) to go to sleep earlier, and reduce active time.
SCell Dormancy and Faster Activation	In Rel-15, SCell deactivation to activation by MAC CE takes longer transition time. SCell dormancy in Rel-16 enables shorter switch delay from Scell dormancy to activated state. The dormant BWP is one of the UE's dedicated BWPs configured by network via dedicated RRC signaling, in which, UE stops monitoring PDCCH in the SCell, but activities such as CSI measurement/reporting and beam management are not impacted.

[2] MediaTek's Bandwidth Part Adaptation Whitepaper

Figure 17 highlights another two prominent features for VoNR voice quality improvements. These features were first introduced to be studied under eVoLTE framework in Rel-14. However, with the mechanism improvements in later releases, they become applicable for implementations in commercial networks.

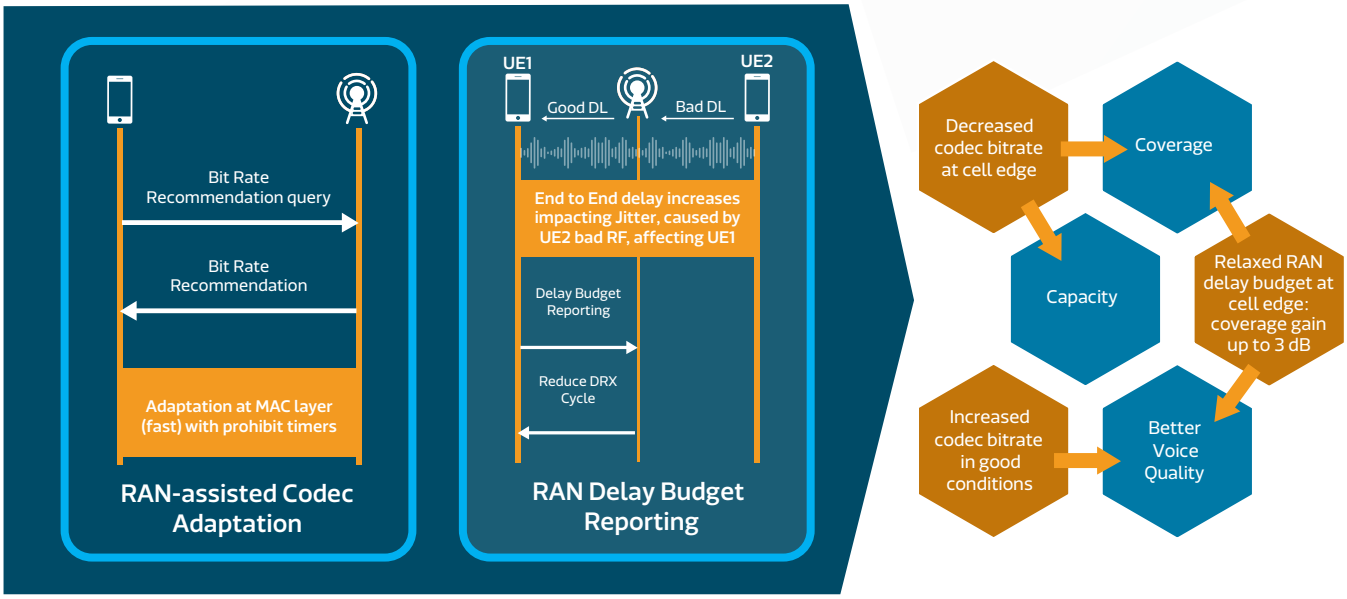


Figure 17. Overview of VoNR Features for Capacity and Coverage Improvements

RAN-assisted codec adaptation (uplink or downlink bit rate adaptation) provides a means for the gNB to send codec adaptation indication with recommended bit rate to assist the UE to select or adapt to a codec rate for voice or video. This mechanism is applicable to uplink/downlink bit rate increase or decrease to improve voice quality in near-cell conditions and the VoNR coverage and capacity. The recommended bit rate for UL and DL is conveyed as a MAC Control Element (CE) from the gNB to the UE, for which gNB may send a recommended bit rate to the UE to inform the UE on the currently recommended transport bit rate on the local uplink or downlink, and UE may initiate an end-to-end bit rate adaptation with its peer (UE or Media Gateway MGW). The recommended bit rate query message is conveyed as a MAC CE from the UE to the gNB, for which the UE may check if a bit rate recommended by its peer can be provided by the gNB. For bit rate recommendation query the value indicates the desired bit rate and as such, the UE is not expected to go beyond the recommended bit rate from the gNB, and shall wait for gNB to take an action based on this query. For bit rate recommendation, the value in MAC CE indicates the recommended bit rate, and conveyed as an index mapped into Kbit/sec bit rate in Table 6.1.3.20-1, 3GPP TS 38.321. The UE advertises its capability of RAN-assisted codec adaptation feature in the UE capability Information message in NR RRC layer with the following IE (information elements):

- recommendedBitRate: Indicates whether the UE supports the bit rate recommendation message from the gNB to the UE.
- recommendedBitRateQuery: Indicates whether the UE supports the bit rate recommendation query message from the UE to the gNB. This field is only applicable if the UE supports recommendedBitRate.
- recommendedBitRateMultiplier-r16: Added in Release-16 for UEs supporting recommended bit rate multiplier, when bitRateMultiplier is configured by gNB. It is intended to support the targeted bitrates are as high as 300 Mbps and beyond, e.g. VR-enabled applications in both the uplink and downlink.

Another feature providing better coverage and voice quality is RAN delay budget reporting. RAN delay budget reporting is based on an RRC framework of UE Assistance Information (UAI) discussed in table 3. Due to delays in VoNR network from different sources, then the voice packets inter-arrival time can vary in time. The typical end-to-end delay (referred to as Mouth-to-Ear delay) is 270 msec which provides a satisfactory voice user experience at the two ends in a voice call. Several sources contribute to this delay including delays at the Uu air interface. A typical Uu delay to maintain a good voice quality is 50 msec. However, such strict delay on the radio side may also affect the coverage when it comes to re-transmissions.

What it means is that the network uses connected mode DRX (cDRX) for VoNR calls with a typical value of 40 msec Long DRX Cycle. However, if two UEs are in a call, and one of which is in bad radio condition, the other UE in good condition may not be aware of the need to relax the DRX cycle in order to give time to the bad UE to perform more re-transmissions and hence improving the coverage. Referring to the mechanism in figure 17, UE1, despite its good coverage conditions, requests and if granted by the gNB, can achieve the shortening of its cDRX cycle, in order to be able to provide more delay budget for UE2 so that UE2 can better tackle its poor coverage conditions and increase the reliability of its uplink transmissions. As such, IMS call quality can be improved through reduced end-to-end delay and jitter. In this flow, UE1 (IMS voice receiver) is in good radio condition and configured with 40 ms cDRX. UE2 (IMS sender) is in bad radio condition and configured with no cDRX. The scenario in figure 17 happens in the following sequence:

- UE2 detects bad-radio condition (e.g., high BLER), it does many HARQ retransmissions, which cause long jitter and E2E delay at the receiver UE1,
- UE1 detects that VoNR quality is bad (e.g., large jitter or delay), hence it suggests gNB1 to de-configure CDRX or shorten CDRX cycle, by sending a *DelayBudgetReport* message to decrease the cDRX cycle length (or even disables it). As a result, end-to-end delay and jitter are reduced,
- UE2 detects that VoNR E2E delay has dropped. UE2 reports larger delay headroom to gNB2, so gNB utilize the additional delay budget to improve the reliability of UE2 uplink transmissions in order to reduce packet loss, e.g., via suitable repetition or retransmission mechanisms.

UEs could estimate the E2E delay budget based on packet loss ratio (PLR), e.g., based on monitoring of RTP receive statistics, RTT thresholds (e.g., with RTT determined by using RTCP sender and receiver reports), or User Plane Latencies. Hence, when both UEs support delay budget report, it becomes best to achieve the desired gain of the feature, while gNB coordination is needed. For example, while an UE receiver in good coverage may turn off cDRX to create delay budget for an UE sender, it may be the case that the UE sender does not even support delay budget reporting, or that the UE sender's gNB may not grant the additional delay budget to the UE sender, so the effort of the UE receiver may not deliver any end-to-end performance gain, and end up wasting the battery power of the IMS receiver UE. The UE advertises its capability of this feature in UE Capability Information message through *delayBudgetReporting*. The network can utilize this feature as part of the overall UAI framework.

Acknowledgements

MediaTek Carrier Engineering Services (CES) is a team working under Wireless System Design and Partnership (WSP) division. CES team has an extensive global experience in modems and utilize it in order to assist mobile network operators in network optimization and strategy planning. CES team is responsible for working with network operators on new technology evaluations and with partners to improve the end-user experience. The work in this paper and other offerings in 5G come from the team's vast experience in 5G deployment scenarios. We wish to express our appreciation to our colleagues in MediaTek R&D and System Design teams who developed extensive studies on EPSFB, VoNR, power saving framework and contributed to several advanced features in 3GPP Release 16 and 17.

Paper Authors:

Deepak Verma

Carrier Engineering Services at MediaTek, Inc. India.

ChinLin Low

Carrier Engineering Services at MediaTek, Inc. Singapore.

Mohamed A. El-saidny

Carrier Engineering Services at MediaTek, Inc. Dubai.