SA WG2 Meeting #S2-120
27 - 31 March 2017, Busan, South Korea

S2-171802
(was S2-17xxxx)

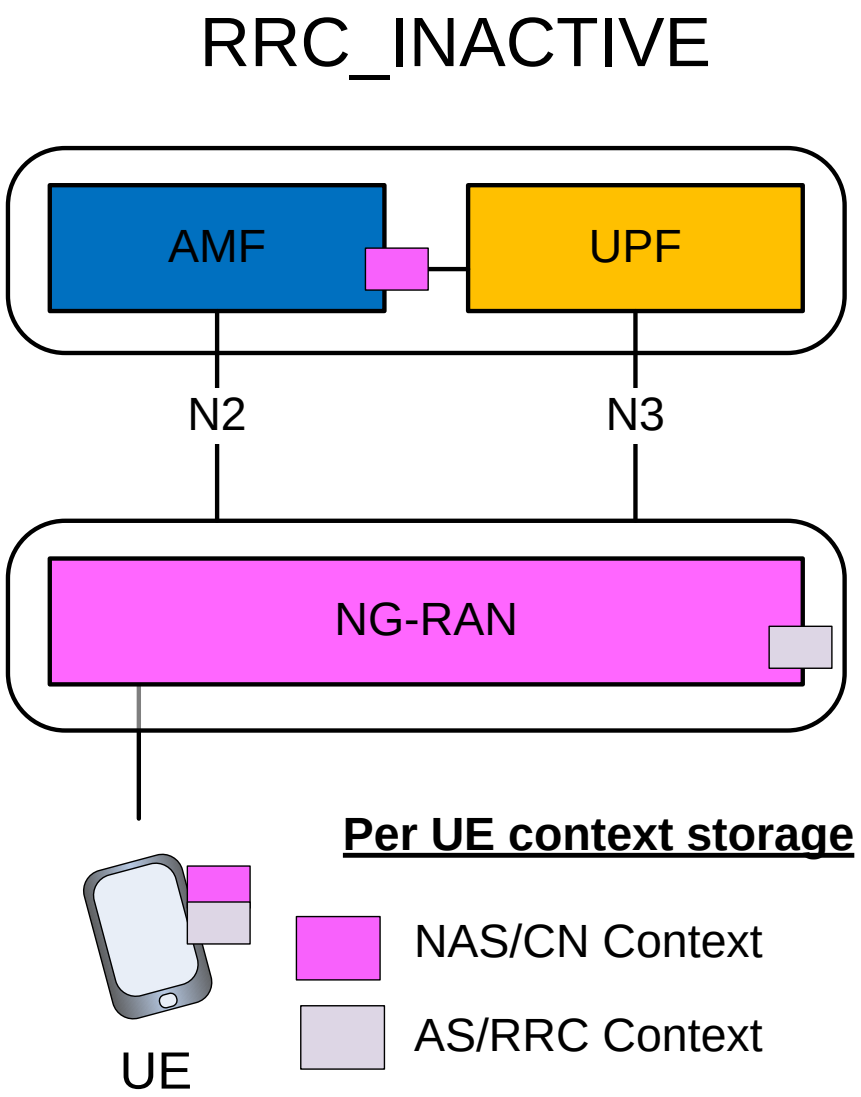
Source: Qualcomm Incorporated
Title: RRC_INACTIVE System architecture aspects
Agenda Item: 6.5.12
Work Item / Release: 5GS_ph1/Rel-15

QUALCOMM®

Background

RRC_INACTIVE State

- RAN: Introduced in RAN WGs 2/3 as new RRC state (RRC_INACTIVE) for NR.
  - Keeps UE context in NR-RAN and N2/N3 interfaces active
  - UE dedicated Resource is released
  - UE performs cell selection/reselection within a RAN-based notification area.
  - Worked under basis that state in NAS/CN is connected
- SA2: There were questions of
  - Why is it needed?
  - What are the system impacts? Many comments that this needs study….
  - Way forward: Dedicated agenda for mini study of RRC_INACTIVE within 5GS_ph1
- Goal of this presentation:
  - Review RRC_INACTIVE objectives as defined by RAN WGs.
  - Understand current RAN2/3 design and justification for it.
  - System architecture impacts and proposed way forward.



Item 1: Motivation for RRC_INACTIVE (1)

- Requirements for RRC_INACTIVE have been captured in TR 38.804 as follows:
  - "Study the introduction of a RAN controlled "state" characterised by, at least:
    - a) UEs in RAN controlled state should incur minimum signalling, minimise power consumption, minimise resource costs in the RAN/CN making it possible to maximise the number of UEs utilising (and benefiting from) this state.
    - b) Able to start data transfer <u>with low delay</u> (as required by RAN requirements)."
      - (From TR 38.913: "For eMBB, the target for user plane latency should be 4ms for UL, and 4ms for DL.")

- A key driver of RAN WGs design has the capability of starting data transfer with low delay at all times.
  - This requires the ability to initiate data transfer without need for round trip delay to Core Network for UE context/security… which requires the UE context to be available in the RAN.
    - **For E-UTRAN/EPC, Qualcomm data shows generally around an order of magnitude difference between signaling across X2 and round trip between eNB and MME via S1.**

- Why a new state (compared to current EPC states/solutions)?
  - Regular **CONNECTED**: Meets the low delay requirement **but does not meet the reduced signaling/power consumption/resource**.
  - Regular **IDLE**: Meets reduced signaling/power consumption requirement, **does not meet the low delay requirement**.
  - **IDLE** with **UP optimization**: **Does not always meet the low delay requirement.** (see example slide later)
  - **IDLE** with **CP optimization: Works only for small data.**
  - **NO current state meets the requirements above for RRC_INACTIVE.**

RAN status: Agreements in TR 38.912 (copy/paste)

General

- **RRC_INACTIVE**:
  - Cell re-selection mobility;
  - CN – NR RAN connection (both C/U-planes) has been established for UE;
  - The UE AS context is stored in at least one gNB and the UE;
  - Paging is initiated by NR RAN;
  - RAN-based notification area is managed by NR RAN;
  - NR RAN knows the RAN-based notification area which the UE belongs to;

Paging

The UE in RRC_IDLE and RRC_INACTIVE states may use Discontinuous Reception (DRX) in order to reduce power consumption.
While in RRC_IDLE the UE monitors CN-initiated paging, in RRC_INACTIVE the UE is reachable via RAN-initiated paging and CN-initiated paging.
RAN and CN paging occasions overlap and same paging mechanism is used. The UE monitors one paging occasion per DRX cycle for the reception of paging as follows:

- Paging DRX cycle length is configurable;
  - A default DRX cycle for CN paging is configurable via system information;
  - A UE specific DRX cycle for CN paging is configurable via UE dedicated signaling;
  - A RAN node can configure a UE with a DRX cycle for RAN paging. This configuration can be UE specific.

RAN notification area

A UE in the RRC_INACTIVE state can be configured with the RAN-based notification area, whereupon:

- a notification area can cover a single or multiple cells, and can be smaller than CN area;
- a UE does not send any "location update" indication when it stays within the boundaries of the notification area;
- leaving the area, a UE updates its location to the network.

There are several different options on how the RAN-based notification area can be configured:

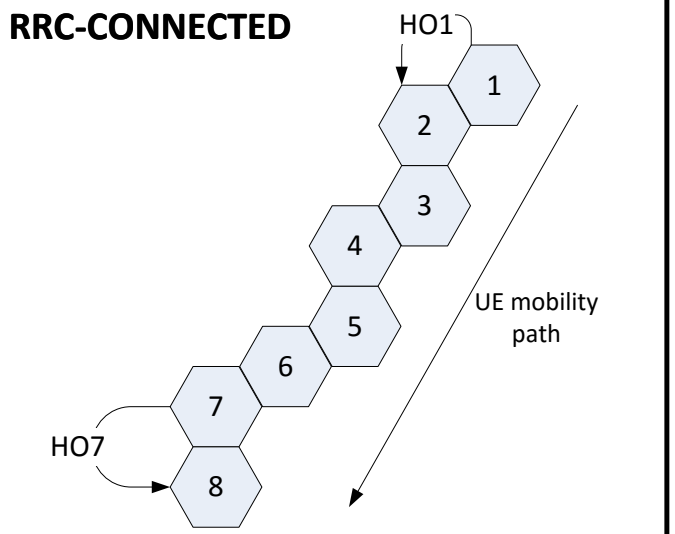
- List of cells;
  - A UE is provided an explicit list of cells (one or more) that constitute the RAN-based notification area.
- RAN area.
  - A UE is provided (at least one) RAN area ID;
  - A cell broadcasts (at least one) RAN area ID in the system information so that a UE knows which area the cell belongs to.

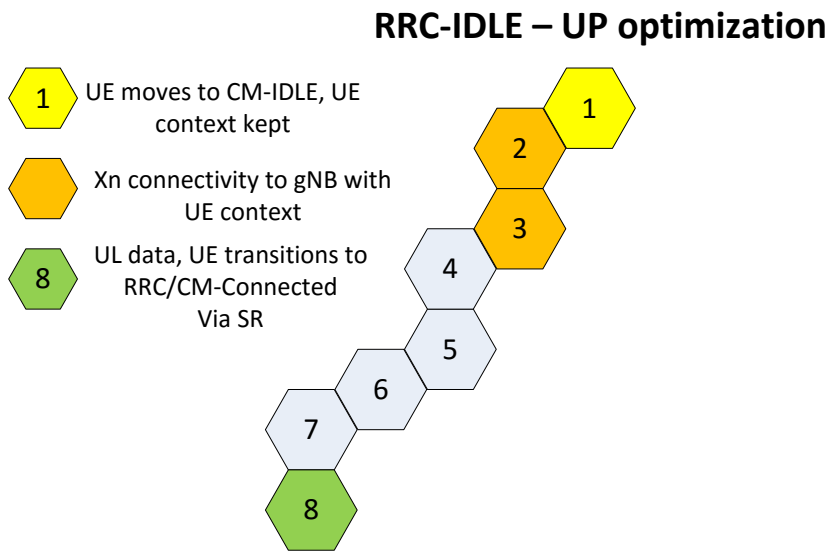
NOTE 1: It will be decided in the work item phase whether to support both options, list of cells and RAN area ID, or only one of them.

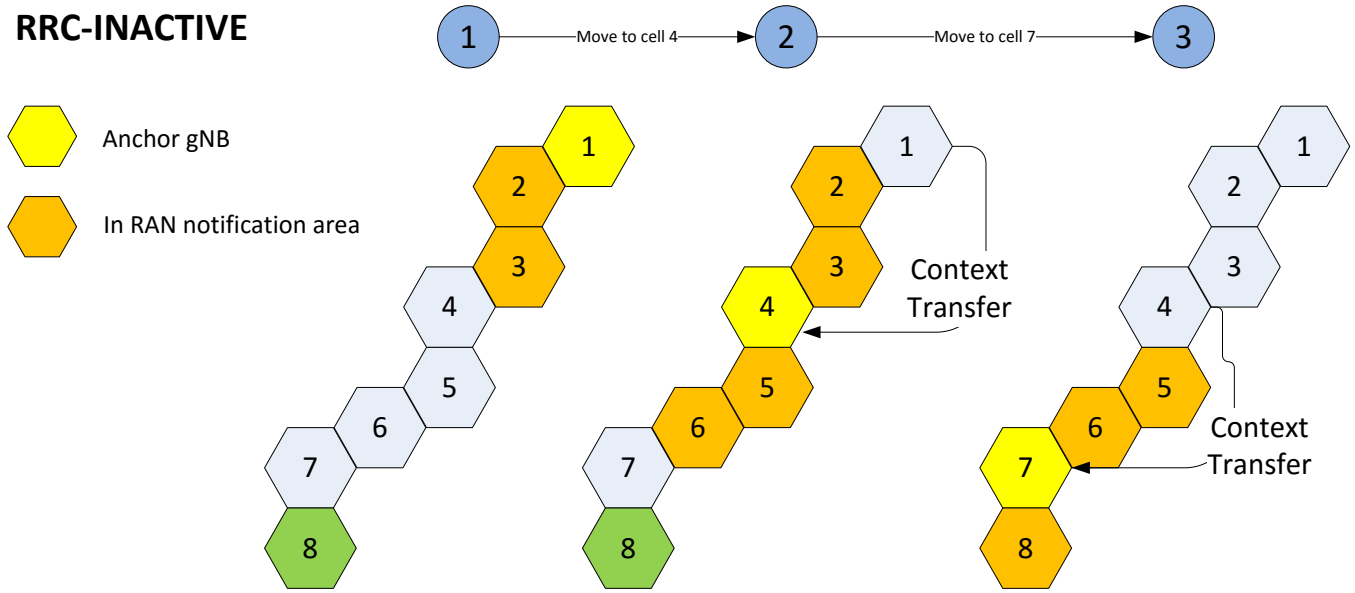
NOTE 2: A list with cells may contain only one entry implementing RAN-based notification area comprising one cell.

Motivation for RAN notification area

- Example:
  - Last data @cell1
  - Moves to cell 8
  - Next UE data @cell8
  - Cells 2/3<->Xn<-> cell 1
- RRC-CONNECTED:
  - 7 HOs before next UE data
  - Context ready @ RAN
  - Best latency, high power/signaling
- RRC_IDLE/UP optimization
  - UE context @ RAN not available from cell 8
  - Full service request
  - Worst latency
- RRC_INACTIVE
  - 2 context transfers when outside RAN notification area
  - Context reachable @ gNB with cell 7.
  - Latency OK always!







System Aspects

- Key questions:
  - **What is NAS Connection Management (CM) state?**
    - RAN2/3 has been assuming CM-CONNECTED, this has also been agreed in study phase in SA2.
    - It seems clear starting point, but these questions might affect the conclusion:
    - **1) What is N2/N3 status?**
    - **2) What is Mobility Model?**
    - **3) Does CN / NAS need to be aware of RRC_INACTIVE state?**
    - **4) Do we need Core Network paging in RRC_INACTIVE?**
  - **Other aspects:**
    - **5) Core network assistance to RAN for RRC_INACTIVE?**
    - **6) How is RRC_INACTIVE enabled/disabled?**
    - **7) Mismatch of states at CN and UE**
    - **8) PLMN selection in RRC_INACTIVE**
    - **9) Inter-RAT mobility between NR and eLTE. Relation to Light Connection?**
    - **10) Inter-RAT mobility in all other cases to GERAN/UTRAN.**
    - **11) UE location accuracy**
    - **12) Access control for RRC_INACTIVE.**

1) N2/N3 status while in RRC_INACTIVE

- Clearly the Core Network must know the gNB that has the UE AS context, so there must be some context towards the anchor gNB, but does it need to be active?:
  - As example, UP optimization keeps S1-C (N2 in 5G) “suspended”, but the key part is S1-U (N3) is not active, and UE must always be CN paged.

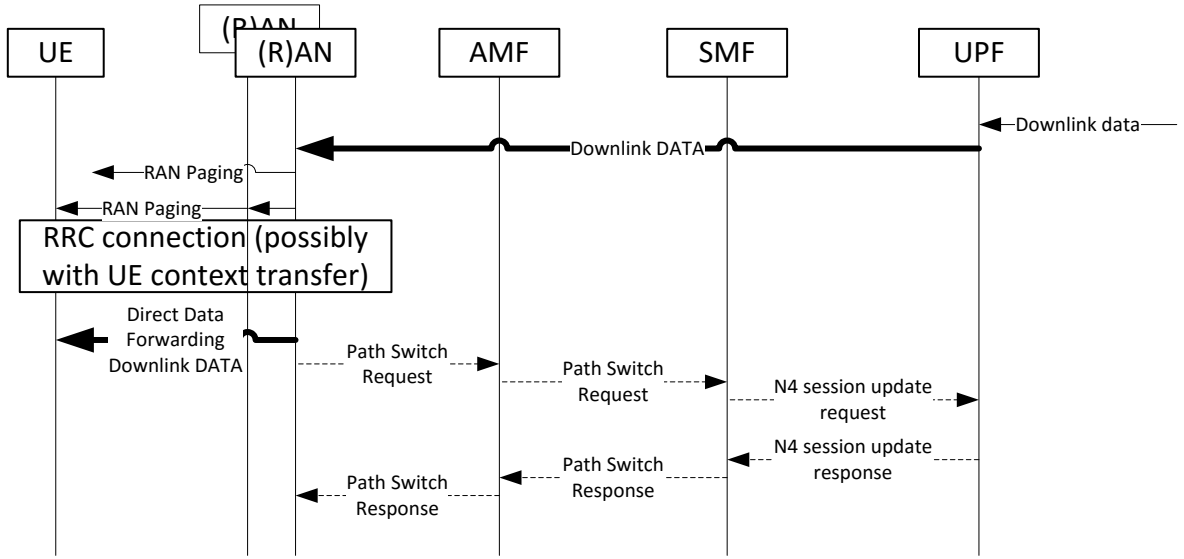
- Since delay is the driving requirement for RRC_INACTIVE, **N3 must be active**, and data must be sent from UPF to gNB directly, so that when UE responds to RAN paging, the data is ready at the RAN (directly or via Xn data forwarding).

- Since the UE always updates the NW when the UE moves outside the RAN notification, the UE is always assumed reachable when the UE is in RRC_INACTIVE.

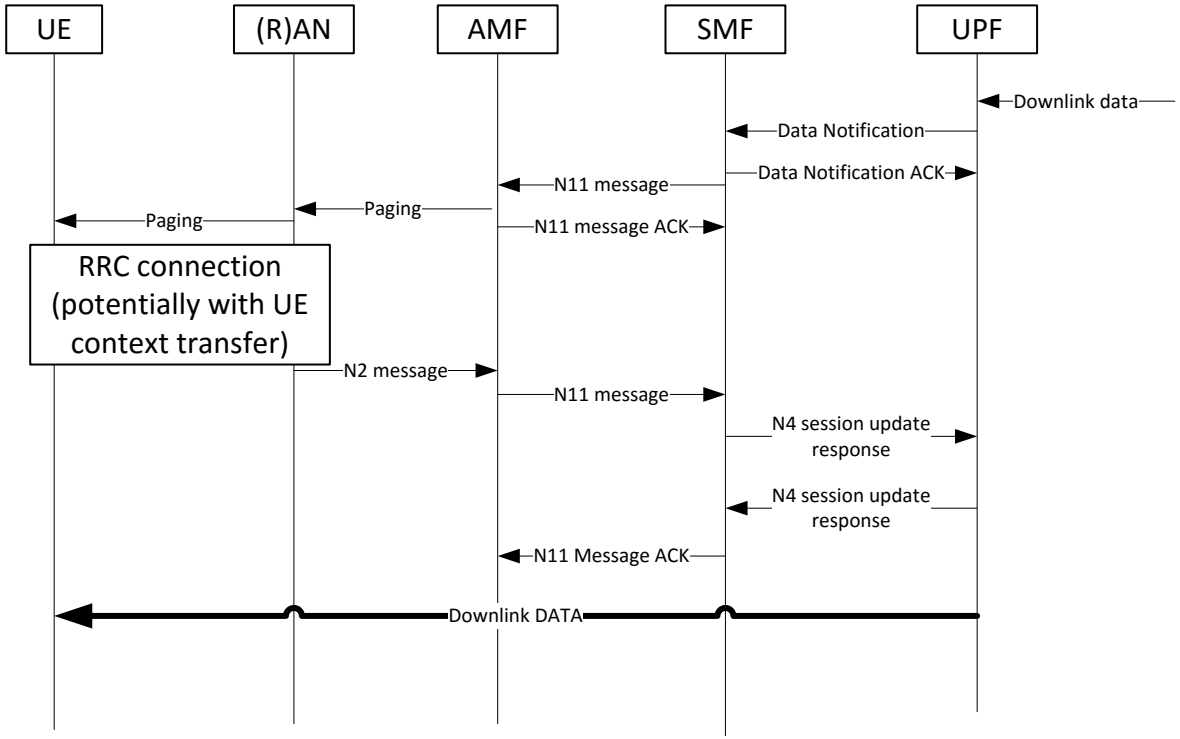
- Since N3 must be active, there’s no reason to not have N2 active as well, (e.g. no reason to have N2 “suspended”).

Latency Comparison for active / not active N3.

- Active N3



- Not active N3



- Active N3 avoids at least whole roundtrip over control plane between UPF <->SMF<-> AMF<->RAN before UPF can send data to RAN.
- With active N3 data is already at RAN when UE responds to RAN paging.

2) Mobility Model (1)

- Driven by:
  - UE context always reachable @ RAN:
    - Anchor gNB always reachable via Xn within the RAN notification area
    - If UE steps out of RAN notification area, UE context must be transferred from anchor gNB to new serving gNB
  - N2/N3 always active.

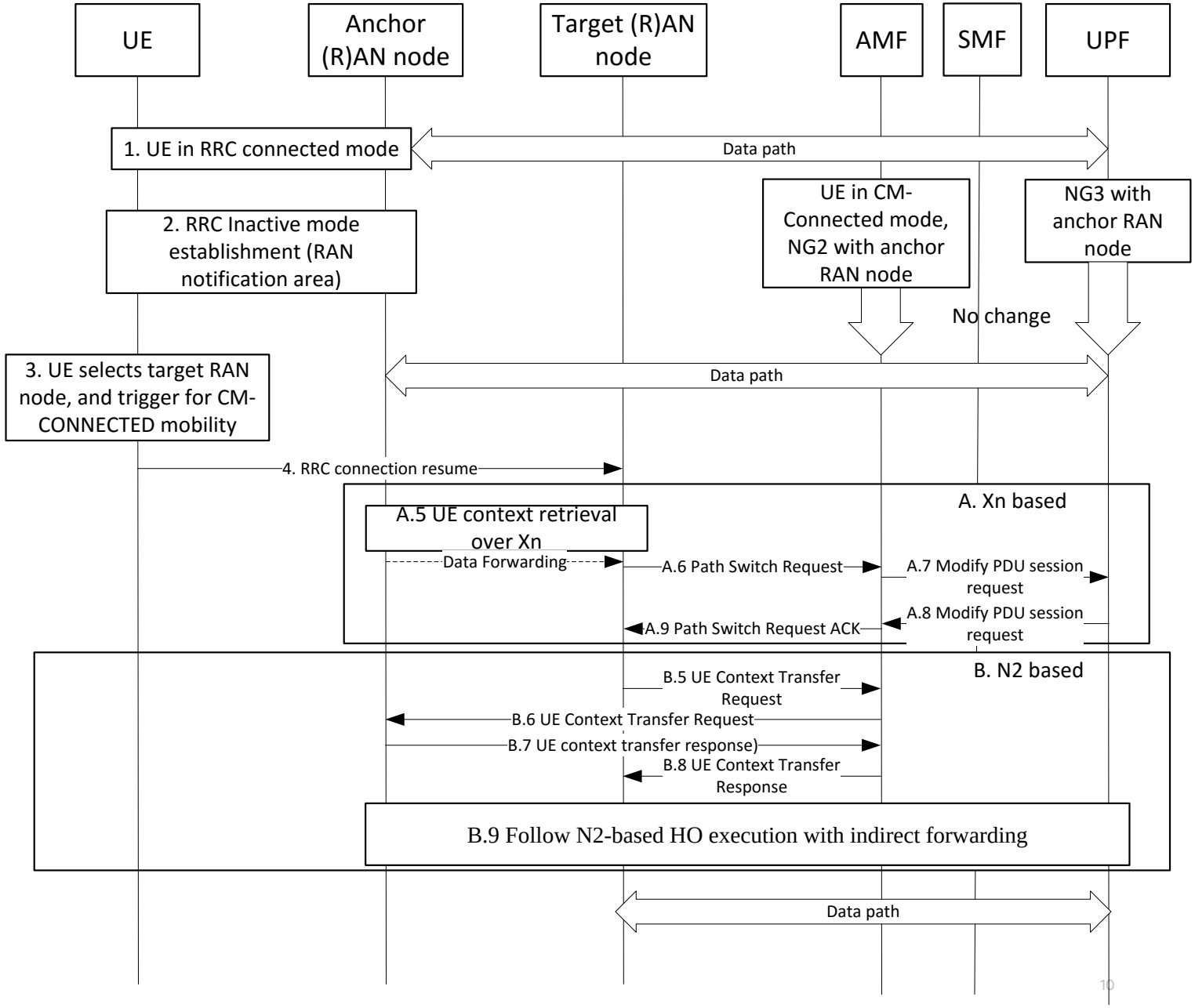
- Triggers for inter-gNB mobility:
  - UE moves outside RAN notification area
  - UE moves from RRC_INACTIVE to RRC_CONNECTED in gNB different from anchor gNB
    - UL data, NAS procedure, as response to RAN paging

- From Core Network perspective:
  - Any trigger for UE context transfer from anchor gNB (original N2/N3 termination) to a new serving gNB (new N2/N3 termination), looks like a transfer of N2/N3 termination point.
    - Exactly the same as a handover procedure.
  - UE context transfer from one gNB to another should reuse same procedures towards the Core Network for both RRC_INACTIVE and RRC_CONNECTED.

2) Mobility Model (2)

- Two options:
  - Xn-based (required within RAN notification area)
  - N2 based (for mobility outside RAN notification area)

- Notes:
  - Call flow does not show RAN procedure details.
  - During N2 path switch procedure (Xn based) or N2 based HO, the target gNB may indicate to CN the UE state (RRC_CONNECTED or RRC_INACTIVE)



3) Does CN / NAS need to be aware of RRC_INACTIVE state?

- There may be the following reasons for the CN to be aware of when the UE is in RRC_INACTIVE
  - Location services:
    - If the AMF knows the UE is in RRC_INACTIVE is may provide an “age” of the location, i.e. last time the UE location was known at cell level.
    - This will depend on how location services are defined in 5G.
  - UE NAS must know in case we define PLMN selection with RRC_INACTIVE (currently PLMN selection occurs only in EMM-IDLE), and potentially for other error cases.

- Note however that the above would not create major system impacts, therefore even if the CN must know whether the UE is RRC_INACTIVE, this can be handled as part of CM-CONNECTED, possibly with a substate.

4) Do we need Core Network paging in RRC_INACTIVE?

- RAN paging failure can occur in two different scenarios:
  - Scenario 1: The UE is Out of Coverage.
    - Given RAN paging reliability assumption, this scenario is as infrequent as RLF in RRC connected mode.
    - CN paging does not actually help in this scenario as a UE that is OOC would not receive CN paging either.
    - We propose that gNB will declare RAN paging failure and release the N2 connection moving the UE to ECM IDLE.
  - Scenario 2: The UE moves outside the RAN notification area at the moment the anchor gNB receives DL data.
    - The UE will immediately initiate an update towards the network.
    - The anchor gNB will perform RAN paging reattempts before declaring RAN paging failure.
      - **Note that while the anchor gNB is performing RAN paging reattempts it is not declaring RAN paging failure yet, with enough reattempts, this should give ample time for the UE and system to recover before failure is declared.** There is though increased delay cost.
    - This will result in either:
      - Xn based UE context transfer and direct data forwarding occurs before anchor gNB declares RAN paging failure
      - N2 based UE context transfer and indirect data forwarding occurs before anchor gNB declares RAN paging failure
- Conclusion: No, it is not needed, since the unrecoverable failure rate should be around same rate as for connected mode RLF.
- NOTE:
  - If there's interest in studying "ultra reliable communication" i.e. reducing data loss further, this should be a separate study for both RRC_CONNECTED and RRC_INACTIVE (i.e. a new SID would be needed with proper requirements/goals.)

So… do we need a new CM state or can we reuse CM-CONNECTED

- 1) What is N2/N3 status?
  - Conclusion: N2 and N3 must be active and terminating @ anchor gNB. Same as for CM-CONNECTED
- 2) What is Mobility Model?
  - Conclusion: Mobility model from CN point of view can be common with RRC-CONNECTED mobility.
- 3) Does CN / NAS need to be aware of RRC_INACTIVE state?
  - It might in some cases (still FFS) but highly unlikely it results in major change in CN behavior.
- 4) Do we need Core Network paging in RRC_INACTIVE?
  - Conclusion: No.

- Conclusion:
  - CM-CONNECTED state is best state to use in NAS/CM for RRC_INACTIVE.
  - RRC_INACTIVE awareness and slight behavior differences may be needed. However, this does not warrant a new state.

5) Core network assistance to RAN for RRC_INACTIVE?

- RAN needs to receive at least the following information from the AMF before transitioning a UE from RRC_CONNECTED to RRC_INACTIVE:
  - The registration area (TAI list) provided to the UE
    - The anchor gNB shall assign a RAN notification area that is contained by the registration area.
    - The RAN notification area can cross tracking areas as long as they are in the TAI list.
  - A maximum allowed DRX value
    - This could be the UE specific I-DRX, the DRX used for RRC-INACTIVE shall always be below this value.
    - If we define UE provided preference @ NAS, this could be based on UE provided information of maximum delay tolerance.
  - Possibly:
    - Some prioritization information (this needs further study, might be useful if there's RAN paging prioritization in 5G CN)

- This information is provided during N2 activation with the (new) anchor gNB (attach, service request, handover).

6) Should CN have control to enable/disable RRC_INACTIVE?

- The AMF can implicit indicate to RAN RRC_INACTIVE is enabled by providing the RRC_INACTIVE assistance information (lack of this means not enabled).
  - RAN actually decides if and when to transition UE to RRC_INACTIVE if AMF enabled it.

- The AMF may decide to enable RRC_INACTIVE:
  - Based on subscription information
  - Based on UE provided latency requirements
    - The UE may actually request the use of RRC_INACTIVE directly.
  - Based on accepted network slices.
  - Etc.

- The AMF may decide to not enable RRC_INACTIVE:
  - For low power consumption scenarios (e.g. PSM/eDRX equivalent in 5G).
  - Cellular IOT (Rel-16 transfer).
  - Depending on UE mobility pattern (low mobility/stationary) and/or data pattern (infrequent).
    - E.g. UE is moving fast on average and relatively infrequent data.

7) Mismatch of states at CN and UE

- UE in CM-IDLE and AMF in CM-CONNECTED/RRC_INACTIVE
  - When it can occur:
    - UE camps on cell not indicating support of RRC_INACTIVE
    - RRC resume procedure fails
    - Any other failure scenario that cannot be solved within RRC (e.g. prolonged OOC)
  - What happens next:
    - When UE transitions to CM-IDLE, it immediately start NAS recovery, i.e. it either performs Service Request or Registration update procedure.
  - Potential Issues:
    - Data may be lost between time UE detects failure and recovery.
  - Does it warrant a redesign?
    - No, this should be infrequent enough, particularly for homogeneous support of RRC_INACTIVE in large areas.

- UE in CM-CONNECTED/RRC_INACTIVE and AMF in CM-IDLE
  - When it can occur:
    - AMF reset
    - RAN paging failure
  - Solution: UE monitors both RAN and CN paging. If UE receives CN paging it moves to CM-IDLE and replies with either SR or registration update.

8) PLMN selection in RRC_INACTIVE

- Background:
  - PLMN selection is controlled by NAS.
  - Currently performed only in IDLE mode.

- Proposed Solution:
  - NAS is aware of RRC_INACTIVE state.
  - UE configuration for PLMN selection is the same for RRC_IDLE and RRC_INACTIVE.
    - Same list of PLMN priorities
    - Same timers for PLMN selection triggering
  - NAS requests AS to provide list of available PLMNs.
  - If higher priority PLMN is found, NAS moves to CM-IDLE, and continues PLMN selection procedure as per CM-IDLE state.
  - Further actions from UE when it moves to CM-IDLE:
    - To gracefully leave the PLMN and remove the context in the source PLMN RAN, the UE may send an RRC release request (possibly with NAS detach) from RRC_INACTIVE.

9) Mobility between NR and eLTE

- Background:
  - RAN2 has progressed Light Connection solution for E-UTRAN.
    - SA2 will now study the core network aspects for EPC (new SID approved at plenary)
  - Motivation for light connection and RRC_INACTIVE are the same
  - The RAN design has been kept essentially equivalent for light connection and RRC_INACTIVE
  - RAN2 has left FFS whether it is possible to move across RRC_INACTIVE and Light Connection.
- Goal:
  - From System Architecture perspective, we should provide a common core network solution for NR/RRC_INACTIVE and eLTE/Light Connection to facilitate mobility between NR and eLTE without need to move to idle mode or connected mode and update the network, i.e. translating state from NR/RRC_INACTIVE to eLTE/Light Connection.
  - Motivation:
    - 5G CN provides core network functionality to both NR RAN and eLTE.
    - It is likely deployments will be mixed at least in the near future.
    - In general, feature incompatibilities across NR RAN and eLTE will be extremely detrimental to the feature.
    - In particular, allowing the UE to move between NR RAN/eLTE allows to create larger RAN notification areas, decreasing signalling.
    - As example, for idle mode: tracking area lists can contain both NR.
- Proposal:
  - Provide a common 5G CN solution for NR/RRC_INACTIVE and eLTE/Light Connection
  - This also has a RAN component that will need to be studied/adopted in RAN2/3: Light Connection in the context of eLTE.

10) Inter-RAT mobility to GERAN/UTRAN.

- Qualcomm proposal:
  - If a UE in RRC_INACTIVE performs cell selection to GERAN/UTRAN, it shall enter PMM-IDLE state or GPRS STANDBY state
  - Then the UE follows regular idle mode procedures.

For Further Study

- 11) Location services
  - Location services for 5G CN would need to progress further before we can study how it would work with RRC_INACTIVE

- 12) Access Control
  - RAN2 agreed to have a common framework for Access Control for RRC_CONNECTED/RRC_IDLE/RRC_INACTIVE, but no details have been agreed.
  - From TR 38.912:

    One unified access barring mechanism for NR should be introduced to address all the use cases and scenarios that LTE addressed with different specialized mechanisms. The unified access barring mechanism should be forward compatible in order to cope with future use cases/scenarios.

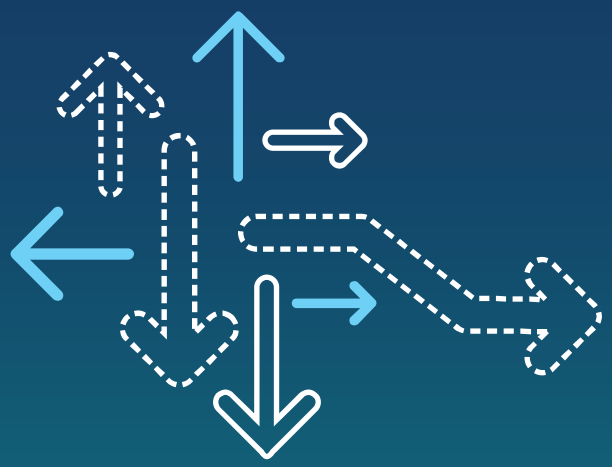
    In NR, the unified access barring mechanism should be applicable for all RRC states in NR (RRC_IDLE, RRC_CONNECTED and RRC_INACTIVE).

    NOTE: It is FFS whether it will be possible for the mechanism to be completely common between the states.

    NOTE: It is FFS if it is possible to specify the unified access barring mechanism fully inside the 3GPP WGs.

  - From SA2 perspective, we prefer RAN2 advances its design and keeps SA2 informed.

ANNEX



Further comparison between RRC_INACTIVE & UP optmization

- Key difference between RRC_INACTIVE and UP optimization design drivers is:
  - **RRC_INACTIVE design** is driven by **guaranteeing UE context always accessible at RAN**:
    - Either Serving gNB is the anchor eNB or anchor gNB is accessible via Xn.
    - **UE must know when UE context is not RAN accessible anymore (by RAN notification area), must update network so UE context is transferred and new RAN notification area is assigned.**
    - **Latency requirement makes it more suitable to keep N3 active, especially for MT data.**
    - Better suited for smartphones with strict delay requirements.
    - Pre Light Connection/RRC_INACTIVE only state that guarantees UE context always accessible at RAN was RRC/CM connected:
      - RRC/CM IDLE can never guarantee this, even with UP optimization.
      - RRC_INACTIVE can be seen as an optimization from RRC/CM_CONNECTED by relaxing the mobility requirements.
        - When discussing signaling optimization/power consumption reduction is by using RRC/CM connected as basis.
  - **UP optimization** is driven by signaling reduction and power consumption, i.e. **Cellular IOT requirements**
    - Latency is not a driving characteristic.
    - Better suited for stationary devices, where most likely UE context remains.
      - Mobile IOT is better covered with Control Plane optimization (data over NAS).
    - It makes sense that UP optimization is an optimization from RRC/CM-IDLE in the context of IOT.

Thank you

Follow us on:

For more information on Qualcomm, visit us at:
www.qualcomm.com & www.qualcomm.com/blog

© 2013 QUALCOMM Incorporated and/or its subsidiaries. All Rights Reserved.

Qualcomm is a trademark of Qualcomm Incorporated, registered in the United States and other countries.
Other products and brand names may be trademarks or registered trademarks of their respective owners.

References in this presentation to “Qualcomm” may mean Qualcomm Incorporated, Qualcomm Technologies, Inc., and/or other subsidiaries or business units within the Qualcomm corporate structure, as applicable.

Qualcomm Incorporated includes Qualcomm’s licensing business, QTL, and the vast majority of its patent portfolio. Qualcomm Technologies, Inc., a wholly-owned subsidiary of Qualcomm Incorporated, operates, along with its subsidiaries, substantially all of Qualcomm’s engineering, research and development functions, and substantially all of its product and services businesses, including its semiconductor business, QCT.