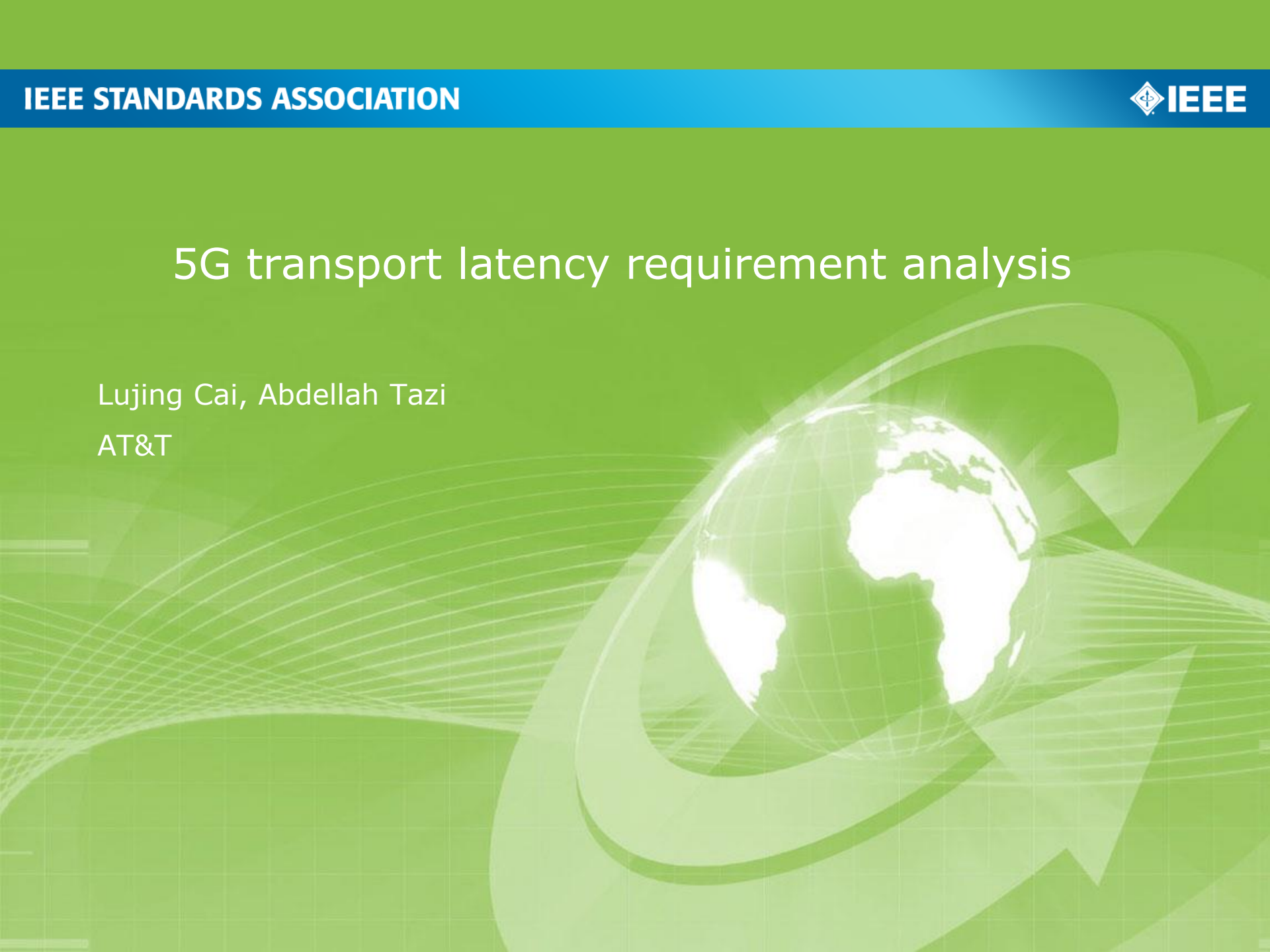
5G transport latency requirement analysis

Lujing Cai, Abdellah Tazi

AT&T



Compliance with IEEE Standards Policies and Procedures

Subclause 5.2.1 of the *IEEE-SA Standards Board Bylaws* states, "While participating in IEEE standards development activities, all participants…shall act in accordance with all applicable laws (nation-based and international), the IEEE Code of Ethics, and with IEEE Standards policies and procedures."

The contributor acknowledges and accepts that this contribution is subject to

- The IEEE Standards copyright policy as stated in the *IEEE-SA Standards Board Bylaws*, section 7, http://standards.ieee.org/develop/policies/bylaws/sect6-7.html#7, and the *IEEE-SA Standards Board Operations Manual*, section 6.1, http://standards.ieee.org/develop/policies/opman/sect6.html
- The IEEE Standards patent policy as stated in the *IEEE-SA Standards Board Bylaws*, section 6, http://standards.ieee.org/guides/bylaws/sect6-7.html#6, and the *IEEE-SA Standards Board Operations Manual*, section 6.3, http://standards.ieee.org/develop/policies/opman/sect6.html

IEEE [WG Project #]
[WG Name]
[WG Chair Name and Email]

5G transport latency requirement analysis			
Date: 2017-04-19			
Author(s):			
Name	**Affiliation**	**Phone [optional]**	**Email [optional]**
Lujing Cai	AT&T		lc779g@att.com
Abdellah Tazi	AT&T		

IMT-2020 RAN latency requirement*

- User-plane latency
  - defined as the one-way time to deliver an packet from the layer 2/3 SDU ingress point to the layer 2/3 SDU egress point of the radio interface
  - assuming unloaded conditions (i.e., a single user) for small IP packets (e.g., 0 byte payload + IP header), for both downlink and uplink
- Control-plan latency
  - the transition time from a most "battery efficient" state (e.g. Idle state) to the start of continuous data transfer (e.g. Active state)

	eMBB	**URLLC**
User-plane latency requirement	4ms	1ms
Control-plane latency requirement	Minimum: 20ms, encouraged: 10ms	

* Reference: "Minimum requirements related to technical performance for IMT-2020 radio interface(s)", ITU-R M. [IMT-2020.TECH PERF REQ], 22 February 2017

3GPP 5G/NR RAN latency requirement*

- User-plane latency
  - The time it takes to successfully deliver an application layer packet from the layer 2/3 SDU ingress point to the radio protocol layer 2/3 SDU egress point via the radio interface in both uplink and downlink
- Control-plane latency
  - the time to move from a battery efficient state (e.g., IDLE) to start of continuous data transfer (e.g., ACTIVE)

	eMBB	**URLLC**
User-plane latency requirement	4ms	0.5ms**
Control-plane latency requirement	10ms	

* Reference: ” Study on Scenarios and Requirements for Next Generation Access Technologies)”, 3GPP TR38.913 v14.2.0 (2017-03)

** With ultra-reliable requirement (<10^{-6} RLC PDU error), the latency requirement may get relaxed to 1ms

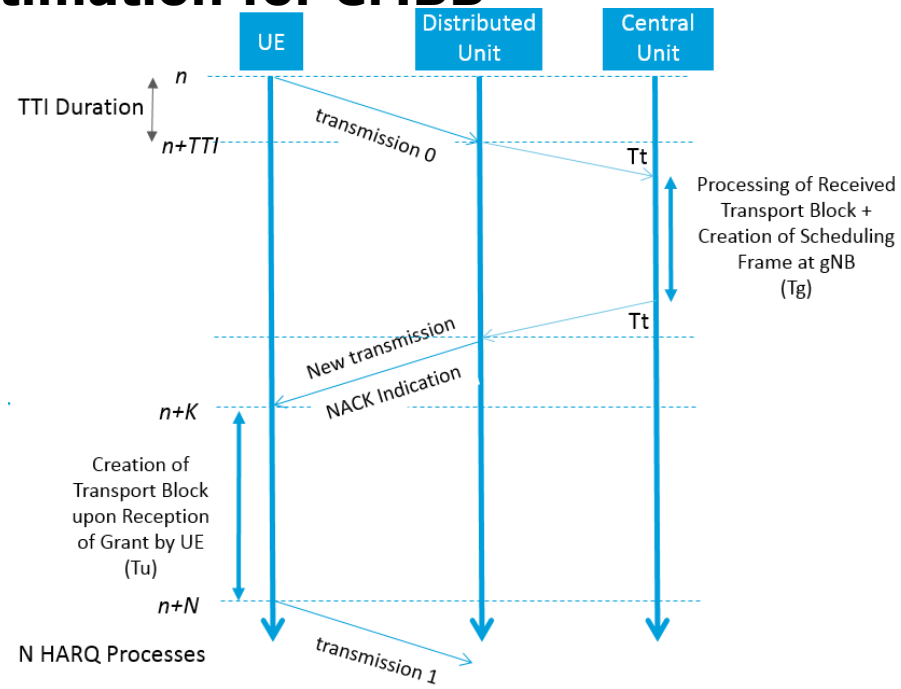
Example of transport latency estimation for eMBB

Interlaced HARQ subframes, low split (option 7/8)

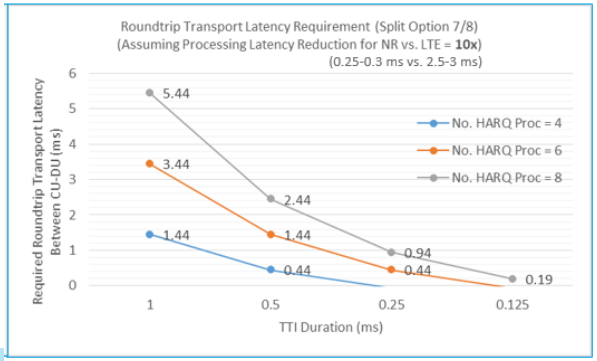
- Transport latency requirement driven by roundtrip HARQ latency, which depends upon:
  - TTI Duration – function of numerology (sub-carrier spacing: … ,60kHz, 120kHz, 240kHz, …)
  - Number of configured HARQ processes
- UE and gNB processing time
- NR design may allow reduction in processing time compared to LTE
  - Front-loaded DMRS design
  - More efficient hardware implementation
- TTI duration of 0.125ms (120/240KHz SCS) is feasible only if transport one way latency <0.1ms

Self-contained HARQ subframes

- Enable the data transmission and corresponding feedback to be contained within the same subframe
- Option 7/8 splits are likely not feasible if not to put too much stress on the transport latency requirement
- Option 2/3-1 functional splits are feasible with similar latency requirements as with interlaced HARQ, i.e., <0.5-2.5ms (one way latency)



Roundtrip Transport Latency = 2*Tt = (N-2)*TTI – Tg – Tu



TTI duration = 0.125 ms feasible with HARQ interlace = 8 and 0.19 ms roundtrip transport latency

Example of transport latency estimation for URLLC

- URLLC service latency requirement
  - 0.5 ms user-plane RAN latency
  - 1 ms end-to-end latency
- Grant-free transmission with automatic and ACK/NACK-less retransmission assumed to meet the target
- Transport latency requirement for Option 7/8 split may be very tight and likely not feasible if not put too much stress on the transport
- Transport latency requirement for Option 2/3-1 split may need to be < 0.05-0.1 ms one way latency in order to meet end-to-end 1 ms service budget

Summary of latency estimates for eMBB & URLLC

All latency estimates are one way	eMBB	URLLC	Comments
OPTION 7-1 or OPTION 7-2	**< 0.1 ms (self-contained subframes not feasible)**	**Likely Not Feasible**	Support of 0.125 ms TTI duration may require gNB/UE processing time = 0.25-0.3 ms and HARQ interlace = 8
OPTION 2 or OPTION 3-1	**0.5 to 2.5ms***	**< 0.05-0.1 ms**	*Further confirmation needed for the eMMB latency estimate

The estimate is preliminary and may need get refined over time

Challenges for accurate latency requirement at this point

- Wide range of TTI durations possible in NR for different combinations of subcarrier spacing, slot duration and level of slot aggregation. Transport latency requirements for any functional split at the PHY or lower L2 layers, would significantly vary in proportion to the slot/TTI durations.
- New HARQ subframe structures potentially adopted in 3GPP (self-contained HARQ, automatic and ACK/NACK-less retransmission, etc.) may have significant impact on the latency requirement for low split options.
- Transport latency requirement depends not only on standards but also on specific equipment implementations
- Function split, a dominant factor impacting transport requirement, has large number of options/sub-options. 3GPP's decision on the function split is not fully clear yet.
- The protocol stack design for NR is not yet stable in 3GPP RAN2. There are possible relocation of some functionalities (reordering, segmentation, etc.) among the stacks, which may impact the latency requirement

Proposed way forward

- First narrow down the COS priority levels according to rank of the latency requirements, i.e., tighter latency data traffic will be assigned to COS with higher priority level. Option to be discussed:
  - Always assign URLLC data traffic to the COS with highest priority level (p0), since this COS also needs the transmission reliability as requirement
  - Assign data traffic of low splits (option 6,7, and 8) to COS with high priority(p1)
  - Assign data traffic of high splits (option 2,3,4, and 5), transport C&M traffic, and RAN-C&M traffic to COS with medium priority (p2)
  - Assign legacy backhaul data traffic to COS with low priority (p3)

- With best effort by now to decide the latency requirement tiers associated to each of the COS priority levels. Options to be discussed for one-way latency requirement:

  [(p0: $\leq 50\mu s$), (p1: $\leq 100\mu s$), (p2: $\leq 1ms$), (p3: $\leq 10ms$)

- To adapt to future requirements when 3GPP decisions become more clear, amendment can be made in another delay requirement profile (e.g. profile 2) if necessary.

Proposed NGFI transport classes of service and KPIs

Class	Sub Class	Priority Level	Latency upper bound requirement (one way)	Throughput requirement (FFS)	Reliability	Reserved	informative
control & management	synchronization	TBD	TBD				
	RAN control-plane	2	τ_1				
data-plane	Subclass_0	0	τ_0		Yes		URLLC Application
	Subclass_1	1	τ_1				3GPP model Option 6,7,8
	Subclass_2	2	τ_2				3GPP model Option 2,3,4,5
	Subclass_3	3	τ_3				Legacy backhaul
Transport NW control & management	Transport NW control-plane	2	τ_2				
Reserved							

	τ_0	τ_1	τ_2	τ_3
Profile 1	$50\mu s$	$100\mu s$	$1ms$	$10ms$
Profile 2	TBD			

$$\tau_0 \leq \tau_1 \leq \tau_2 \leq \tau_3$$

Motion #1

- For COS KPI specifiction, agree to the WF strategies and COS table proposed in slide 9&10 of tf1_1704_cai_tazi_5G_transport_latency_requirement_analysis_1.pdf
- Mover: Lujing Cai
- Seconder:

- Yes: ___ No: ___ Abstain: ___ (technical motion needs >= 2/3)