Gap analysis for URLLC services in 5GC

Hui Ni
Huawei

URLLC KPI requirements (TS 22.261)

Scenario	End-to-end latency (note 3)	Jitter	Survival time	Communication service availability (note 4)	Reliability (note 4)	User experienced data rate	Payload size (note 5)	Traffic density (note 6)	Connection density (note 7)	Service area dimension (note 8)
Discrete automation – motion control (note 1)	1 ms	1 µs	0 ms	99,9999%	99,9999%	1 Mbps up to 10 Mbps	Small	1 Tbps/km^2	100 000/km^2	100 x 100 x 30 m
Discrete automation	10 ms	100 µs	0 ms	99,99%	99,99%	10 Mbps	Small to big	1 Tbps/km^2	100 000/km^2	1000 x 1000 x 30 m
Process automation – remote control	50 ms	20 ms	100 ms	99,9999%	99,9999%	1 Mbps up to 100 Mbps	Small to big	100 Gbps/km^2	1 000/km^2	300 x 300 x 50 m
Process automation – monitoring	50 ms	20 ms	100 ms	99,9%	99,9%	1 Mbps	Small	10 Gbps/km^2	10 000/km^2	300 x 300 x 50
Electricity distribution – medium voltage	25 ms	25 ms	25 ms	99,9%	99,9%	10 Mbps	Small to big	10 Gbps/km^2	1 000/km^2	100 km along power line
Electricity distribution – high voltage (note 2)	5 ms	1 ms	10 ms	99,9999%	99,9999%	10 Mbps	Small	100 Gbps/km^2	1 000/km^2 (note 9)	200 km along power line
Intelligent transport systems – infrastructure backhaul	10 ms	20 ms	100 ms	99,9999%	99,9999%	10 Mbps	Small to big	10 Gbps/km^2	1 000/km^2	2 km along a road
Tactile interaction (note 1)	0,5 ms	TBC	TBC	[99,999%]	[99,999%]	[Low]	[Small]	[Low]	[Low]	TBC
Remote control	[5 ms]	TBC	TBC	[99,999%]	[99,999%]	[From low to 10 Mbps]	[Small to big]	[Low]	[Low]	TBC

NOTE 1: Traffic prioritization and hosting services close to the end-user may be helpful in reaching the lowest latency values.
NOTE 2: Currently realised via wired communication lines.
NOTE 3: This is the end-to-end latency the service requires. The end-to-end latency is not completely allocated to the 5G system in case other networks are in the communication path.
NOTE 4: Communication service availability relates to the service interfaces, reliability relates to a given node. Reliability should be equal or higher than communication service availability.
NOTE 5: Small: payload typically ≤ 256 bytes
NOTE 6: Based on the assumption that all connected applications within the service volume require the user experienced data rate.
NOTE 7: Under the assumption of 100% 5G penetration.
NOTE 8: Estimates of maximum dimensions; the last figure is the vertical dimension.
NOTE 9: In dense urban areas.
NOTE 10: All the values in this table are targeted values and not strict requirements.

Note: requirements on 1ms e2e latency and 99.9999% reliability are under discussion in SA1, this discussion don't depends on the values under discussion.

Factors causing latency/jitter need to be investigated to guarantee URLLC KPIs

- → Latencies/Jitter caused by internal mechanisms within transmission nodes (switch, router, UPF) are typically on μs level ($x \sim xxx\ \mu s$).
- → Usually for traffics of normal latency (e.g. 50-100 ms), the above internal latencies/jitter can be tolerable or ignorable.
- → For service of ultra-low latency on $x(ms)$ level with jitter of μs level, it's necessary to investigate these factors to ensure the stringent KPIs can be guaranteed.
- → The following discussion try to figure out the contributions of different factors to the end-to-end latency and its variation (i.e. the Jitter)

This slides will discuss the factors that may lead to latency/jitter and reliability issues in the core network, and try to estimate their contributions to the KPIs.

Based on that, a way forward is proposed. The SID proposal in S2-180680 may be updated based on the discussion.

Latency: end-to-end link

→ The latency T_{e2e} is composed of

- $T_{propagation}$: Propagation Time on the cable
  - $T_{propagation} = (physical\ cable\ length)/(propagation\ speed)$
  - Considering 40km optical fiber, $T_{propagation} = 40km/(200km/ms) = 200\mu s$
- Sum of T_{device_i}: Latency introduced by each node (UPF, router, switch) on the transmission path
  - Note: Latency caused by RAN is not discussed in this slides.

$$T_{e2e} = T_{propagation} + \sum_{i=1}^{n} T_{device_i}$$

Where T_{device_i} is discussed in next slide

Latency: within one node

→ The latency within one node T_{device_i} is composed of

- $T_{process}$: Time for processing the packet (e.g. tunnel encapsulation/decapsulation, checksum verification, encryption/decryption)
  - For L3 device $T_{process} \approx 200\mu s$
  - For switch of Gbps $T_{process} \approx 1\mu s$
- $T_{transmission}$: Time for transmitting the packet to the link
  - $T_{transmission} = (Packet\ length)/(ReservedOutputBandwidth)$
- $T_{interference}$: Time caused by interference
  - Depends on the congestion level and scheduling mechanism

$$T_{e2e} = T_{propagation} + \sum_{i=1}^{n} (T_{process} + T_{transmission} + T_{interference})_{for\ node\ i}$$

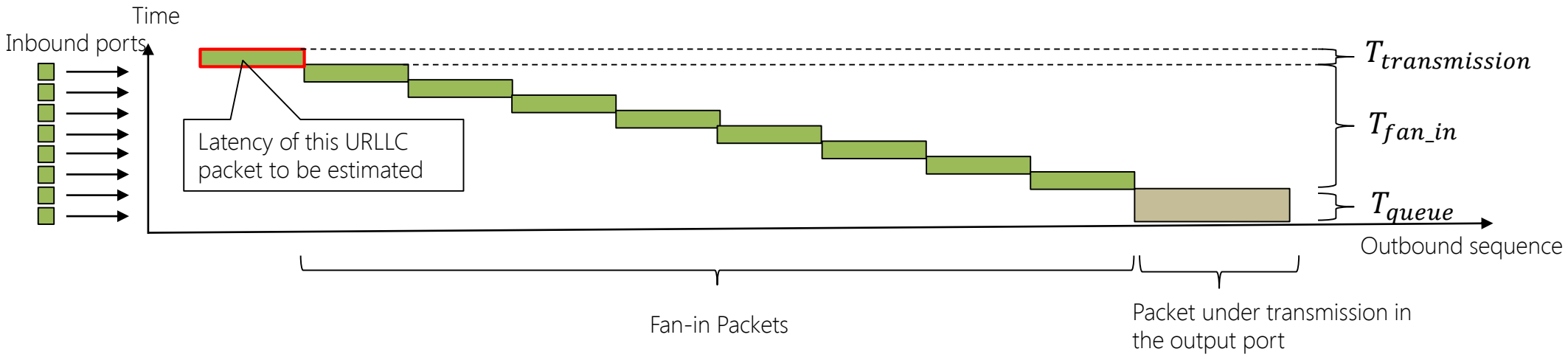
Where $T_{interference}$ is discussed in next slide

Latency: the interference delay

→ For a priority based scheduling mechanism, $T_{interference}$ is composed of

- T_{queue}: Queuing delay
  - The delay caused by the frame under transmission when the URLLC packet arrives, plus the delay caused by queued frames with higher priority than the URLLC packet
- T_{fan_in}: Fan-in delay
  - The delay caused by other URLLC frames arriving the node simultaneously from different input ports within a scheduling interval.
- T_{perm}: Permanent delay
  - The delay caused by other burst packet before the packet. If the reserved bandwidth is greater than the input traffic rate, the permanent delay can be avoid.

Note: If different scheduling mechanisms are adopted, T_{queue} and T_{fan_in} will be different.



Estimation the interference delay for URLLC packet

$$T_{e2e} = T_{propagation} + \sum_{i=1}^{n} (T_{process} + T_{transmission} + T_{queue} + T_{fan_in})_{of\ each\ node\ i}$$

- Assuming URLLC is of the highest priority, $Max(T_{queue}) = maxsizeframe/outputBandwidth$
- $Max(T_{fan_in}) = URLLCpacketLength * (InputPortNum - 1)/ReservedOutputBandwidth$ (*)

  - Note 1: The above calculation assumes the reserved output bandwidth is high enough to avoid congestion in the node. Otherwise once congestion happens, the latency will increase unpredictably.
  - Note 2: The number of fan-in packets depends on traffic pattern and the scheduling within the device.
  - Note 3: if the number of processing pipelines is less than the input port number, processing delay $T_{process}$ will increase due to processor queue (not considered in this discussion).

An example of latency/jitter estimation

- For an example use case with the following assumptions
  - URLLC flow is of highest priority in the network;
  - 50% bandwidth of each node(UPF/router/switch) has been reserved to the URLLC flow; line rate of each port is 1Gbps;
  - Each node has 24 input ports for URLLC flow;
  - URLLC packet length is of 250 bytes, including frame delimiter (SFD) (8 octets), interpacket gap (IPG) (12 octets), and any tunnel header.
- So
  - $Max(T_{queue}) = (1522 + 20) * 8bit/(1Gbps) \approx 12.5\mu s$
  - $Max(T_{fan_in}) = 250 * 8bit * (24 - 1)/(1Gbps * 50\%) \approx 92\mu s$
- For a end-to-end link of 40km cable including 1 UPF and 6 switches
  - $Max(T_{e2e}) = 40km/(0.2km/\mu s) + 200\mu s + 6 * 1\mu s + 250 * 8bit/(1Gbps * 50\%) + 7 * (12.5\mu s + 92\mu s) = 1142\mu s$
  - $Min(T_{e2e}) = 40km/(0.2km/\mu s) + 200\mu s + 6 * 1\mu s + 250 * 8bit/(1Gbps * 50\%) = 410\mu s$
  - $Jitter = Max(T_{e2e}) - Min(T_{e2e}) = 819\mu s$

Observation 1

- → The controllable end-to-end latency depends on end-to-end resource reservation, including transmission network. Congestion must be avoid.
- → Latency/Jitter heavily depends on the scheduling mechanisms used by the nodes on the user plane path.
- → For a fixed transmission path using priority based scheduling, Jitter is mainly introduced by interference packets.
- → Propagation time caused by transmission distance contributes rather a part of end-to-end latency.

High reliability

→ Reliability defined in SA1

- percentage value of the amount of sent network layer packets successfully delivered to a given node within the time constraint required by the targeted service, divided by the total number of sent network layer packets.
- i.e. Reliability == 1- packet loss ratio, according to the above definition
- High reliability depends on an appropriate packet scheduling/transmission mechanism over an network infrastructure (incl. interfaces, connections and functions) that can provide high availability

→ Assuming an appropriate packet scheduling mechanism can always transmit the packet on time, then the “high reliability” will require the “high availability” of network.

High reliability (cont.)

$$Network\ Availability = 1 - \sum_{i=1}^{n}(1 - A_{device_i})$$

$$A_{device} = \frac{Total\ time - DownTime}{Total\ time}$$

Most telco devices are here

Cost	Availability	Average DownTime per Year
	99.9999%	31.6 seconds
	99.999%	5 minutes 16 seconds
	99.995%	26 minutes 18 seconds
	99.95%	4 hours 23 minutes
	99.9%	8 hours 46 minutes
	99.5%	43 hours 50 minutes
	99%	87 hours 40 minutes

→ For an example use case, considering a single transmission path composed of 10 nodes

- To provide a URLLC service of reliability of 99.999% over this path, the average availability of each node cannot lower than 99.9999%, which causes a rapid increasing CAPEX for the deployment.

→ To save the CAPEX, mechanisms have been investigated in the industry to meet high reliability over links with lower reliability. e.g.

- HSR (High reliability Seamless Redundancy) in IEC62439
- PRP (Parallel Redundancy Protocol) in IEC62439
- FRER (Frame Replication and Elimination for Reliability) in IEEE 802.1CB

Note: Mechanisms listed above are just for reference.

Observation 2

→ The mechanism to achieve high reliability over transmission paths of lower availability can save the CAPEX significantly.

Proposed way forward

- → Investigate the impacts of different scheduling mechanisms to latency and jitter;
- → Study mechanisms to minimize the jitter caused by interference packets;
- → Study enhanced mechanisms to shorten the user plane path.
- → Study mechanisms to achieve high reliability over transmission paths with lower reliability.

THANK YOU