

**3rd Generation Partnership Project;
Technical Specification Group Radio Access Network;
Evolved Universal Terrestrial Radio Access (E-UTRA);
Study on latency reduction techniques for LTE
(Release 13)**



The present document has been developed within the 3rd Generation Partnership Project (3GPPTM) and may be further elaborated for the purposes of 3GPP. The present document has not been subject to any approval process by the 3GPP Organizational Partners and shall not be implemented. This Report is provided for future development work within 3GPP only. The Organizational Partners accept no liability for any use of this Specification. Specifications and Reports for implementation of the 3GPPTM system should be obtained via the 3GPP Organizational Partners' Publications Offices.

Keywords

LTE, Radio

3GPP

Postal address

3GPP support office address

650 Route des Lucioles - Sophia Antipolis
Valbonne - FRANCE
Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Internet<http://www.3gpp.org>

Copyright Notification

No part may be reproduced except as authorized by written permission.
The copyright and the foregoing restriction extend to reproduction in all media.

© 2015, 3GPP Organizational Partners (ARIB, ATIS, CCSA, ETSI, TSDSI, TTA, TTC).
All rights reserved.

UMTS™ is a Trade Mark of ETSI registered for the benefit of its members
3GPP™ is a Trade Mark of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners
LTE™ is a Trade Mark of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners
GSM® and the GSM logo are registered and owned by the GSM Association

Contents

Foreword.....	5
Introduction	6
1 Scope.....	6
2 References.....	6
3 Definitions, symbols and abbreviations	7
3.1 Definitions	7
3.2 Symbols	7
3.3 Abbreviations.....	7
4 Study Objectives	7
5 Overview of LTE latency.....	8
5.1 Delay components.....	8
5.1.1 UL and DL latency.....	8
5.1.2 Handover latency [11].....	10
5.2 Current performance	12
5.2.1 UL and DL latency.....	12
5.2.2 Handover latency [11].....	13
5.3 Existing means to limit latency	14
6 Scenarios, Applications and Use Cases.....	14
7 Evaluation Structure and Assumptions	15
8 Solutions for latency reduction	15
8.1 Semi-Persistent Scheduling.....	15
8.2 UL Grant reception	15
8.2.1 Configured SPS activation and deactivation	15
8.3 Handover Latency.....	16
8.3.1 Solution 1: RACH-less handover.....	16
8.3.2 Solution 2: Maintaining Source eNB Connection during Handover	17
8.4 Contention based PUSCH transmission	17
8.4.1 Solution 1 [16]	18
8.4.2 Solution 2 [17]	18
8.5 Reduced TTI and processing time.....	18
9 Performance Evaluation	19
9.1 Protocol evaluations on TTI reduction and Fast UL	19
9.1.1 Simulation 1	19
9.1.2 Simulation 2	22
9.1.3 Simulation 3.....	25
9.1.4 Simulation 4.....	25
9.1.4.1 Simulation assumptions	26
9.1.4.2 TTI shortening simulations	27
9.1.4.3 Fast UL system simulations	29
9.1.5 Simulation 5.....	33
9.1.5.1 General information	33
9.1.5.2 Simulation assumptions and parameters	34
9.1.5.3 Performance results.....	34
9.1.5.3.1 Various backhaul latency	34
9.1.5.3.2 Various FTP file size.....	36
9.1.5.3.3 Various Uu throughput.....	38
9.1.5.3.4 Fast UL access	40
9.1.6 Simulation 6.....	42
9.1.6.1 Simulation assumptions	42
9.1.7 Simulation 7.....	44
9.1.7.1 Simulation setup.....	44
9.1.7.2 Simulated schemes.....	45

9.1.7.3	Simulation results for shorter TTI.....	46
9.1.7.4	Simulation results for Fast UL grant with shorter TTI.....	48
9.1.8	Simulation 8.....	50
9.1.8.1	Simulation assumptions	50
9.1.8.2	Evaluation results.....	51
9.1.9	Simulation 9.....	52
9.1.9.1	Simulation assumptions	53
9.1.9.1	Evaluation results.....	53
9.1.10	Simulation 10.....	56
9.1.10.1	Faster UE Feedback and Rate Control	56
9.1.10.2	Simulation Assumptions:	57
9.1.10.3	Simulation Results	57
9.1.11	Simulation 11.....	58
9.1.11.1	Simulation assumptions	58
9.1.11.2	Evaluation results.....	58
9.1.12	Simulation 12.....	59
9.1.12.1	Simulation assumptions	59
9.1.12.2	Evaluation results.....	60
9.2	Protocol evaluations on Contention based PUSCH transmission	62
9.2.1	Evaluation 1 on solution 1 [16].....	62
9.2.1.1	Resource efficiency.....	62
9.2.1.2	Uplink latency.....	62
9.2.2	Evaluation on solution 2 [17].....	64
9.2.2.1	Resource efficiency analysis on current solutions	64
9.2.2.2	Uplink latency.....	65
9.2.2.3	Resource efficiency.....	66
9.2.3	Evaluation 2 on solution 1 [18].....	67
9.3	Handover latency [11].....	71
9.4	Findings from system evaluations on TTI reduction and reduced processing time	71
9.5	Findings from link evaluations on TTI reduction and reduced processing time	71
10	Conclusion	71
10.1	RAN2 Protocol Evaluations.....	71
Annex A:	Simulation assumptions	73
A1	Protocol Simulations	73
A1.1	Simulation 1 [4]	73
A1.2	Simulation 2 [5]	75
A1.3	Simulation 4 [6]	76
A1.4	Simulation 9 [6]	77
A1.5	Simulation 5 [6]	79
A1.6	Simulation 11 [9]	81
A1.7	Simulation on contention based PUSCH transmission	84
A2	System simulation assumptions for reduced TTI and processing delay	86
A3	Link level simulation assumptions.....	87
Annex B:	System evaluation results.....	91
Annex C:	Link-level evaluation results	91

Foreword

This Technical Report has been produced by the 3rd Generation Partnership Project (3GPP).

The contents of the present document are subject to continuing work within the TSG and may change following formal TSG approval. Should the TSG modify the contents of the present document, it will be re-released by the TSG with an identifying change of release date and an increase in version number as follows:

Version x.y.z

where:

- x the first digit:
 - 1 presented to TSG for information;
 - 2 presented to TSG for approval;
 - 3 or greater indicates TSG approved document under change control.
- y the second digit is incremented for all changes of substance, i.e. technical enhancements, corrections, updates, etc.
- z the third digit is incremented when editorial only changes have been incorporated in the document.

Introduction

1 Scope

This document is related to the technical report for the study item “Study on latency reduction techniques for LTE” [2]. The purpose of this TR is to capture the findings from TSG RAN WG2 and WG1 according to their respective objectives, and to draw a conclusion on way forward.

This activity involves the Radio Access work area of the 3GPP studies and has potential impacts both on the Mobile Equipment and Access Network of the 3GPP systems.

This document is a ‘living’ document, i.e. it is permanently updated and presented to TSG-RAN meetings.

2 References

The following documents contain provisions, which through reference in this text constitute provisions of the present document.

- References are either specific (identified by date of publication, edition number, version number, etc.) or non-specific.
- For a specific reference, subsequent revisions do not apply.
- For a non-specific reference, the latest version applies. In the case of a reference to a 3GPP document (including a GSM document), a non-specific reference implicitly refers to the latest version of that document *in the same Release as the present document*.

- [1] 3GPP TR 21.905: "Vocabulary for 3GPP Specifications".
- [2] RP-150465, New SI proposal: Study on Latency reduction techniques for LTE, Ericsson, RAN#67, March 2015.
- [3] R2-152326, Latency reductions in LTE, Ericsson, RAN2#90, May 2015
- [4] R2-152174, Impact of latency reduction on TCP slow-start behaviour, Intel Corp. RAN2#90, May 2015
- [5] R2-152456, Evaluation on the gains provided by 0.5ms TTI, Huawei, HiSilicon, RAN2#90, May 2015
- [6] R2-154743, Email discussion report for [91#29][LTE/Latency] Evaluation results, RAN2#91, October 2015
- [7] 3GPP TR 36.814: “Further advancements for E-UTRA physical layer aspects”.
- [8] 3GPP TS 36.213: “Physical layer procedures”.
- [9] R2-156340 Performance evaluation on short TTI, ZTE Corporation, RAN2#92, Nov 2015
- [10] R2-156278 L2 overhead for shorter TTI, Nokia Networks, RAN2#92, Nov 2015
- [11] R2-156202, Report of email discussion [91bis#35] [LTE/LATRED] Handover evaluations and solutions, RAN2#92, November 2015
- [12] 3GPP TS 36.300, “Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access (E-UTRAN); Overall description; Stage 2”
- [13] 3GPP TS 36.331, “Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification”

- [14] 3GPP TS 36.133, “Evolved Universal Terrestrial Radio Access (E-UTRA); Requirements for support of radio resource management”
- [15] 3GPP TS 36.321, “Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification”
- [16] R2-154191, Contention based uplink transmission, Huawei, HiSilicon, RAN2#91bis, October 2015
- [17] R2-154122, Analysis on resource efficiency of uplink access solutions, CATT, CATR, RAN2#91bis, October 2015
- [18] R2-156402, Performance evaluation of CB-PUSCH, Nokia Networks, RAN2#92, November 2015

3 Definitions, symbols and abbreviations

3.1 Definitions

For the purposes of the present document, the terms and definitions given in 3GPP TR 21.905 [1] and the following apply. A term defined in the present document takes precedence over the definition of the same term, if any, in 3GPP TR 21.905 [1].

3.2 Symbols

For the purposes of the present document, the following symbols apply:

Symbol format (EW)

<symbol> <Explanation>

3.3 Abbreviations

For the purposes of the present document, the abbreviations given in 3GPP TR 21.905 [1] and the following apply. An abbreviation defined in the present document takes precedence over the definition of the same abbreviation, if any, in 3GPP TR 21.905 [1].

Abbreviation format (EW)

<ACRONYM> <Explanation>

4 Study Objectives

Potential gains like reduced response time and improved TCP throughput due to latency improvements on typical applications and use cases are identified and documented. In this evaluation, latency reductions due to protocol enhancements as well as shortened TTIs are assumed. Specifically, the following items should apply:

- Consider the web application (HTTP/FTP+TCP) use case and analyse the possible gains in the performance metrics of delay and perceived data rate for TCP based data transactions.
- Consider real-time application use case and analyse the possible gains in delay, service coverage and system capacity.
- Other areas in accordance with [2] may be considered as well
- Aspects of complexity, energy-consumption, signalling overhead and resource efficiency should be considered.

The evaluations include active UEs and UEs that have been inactive a longer time, but are kept in RRC Connected. Reducing user plane latency for the scheduled UL transmission with resource efficient solution as a result of protocol and signalling enhancements are studied and compared to the pre-scheduling solutions allowed by the standard today.

Specification impact, feasibility and performance of TTI lengths between 0.5ms and one OFDM symbol are studied, taking into account impact on e.g. L2 overhead, reference signals and physical layer control signalling.

Backwards compatibility shall be preserved, i.e. solutions should allow normal operation of pre-Rel 13 UEs on the same carrier.

5 Overview of LTE latency

In an LTE system there are multiple components contributing to the total end to end delay for connected UEs. The limitations in performance are in general use case dependent; for which e.g. UL latency may influence the DL application performance and vice versa.

5.1 Delay components

5.1.1 UL and DL latency

Examples of sources to latency are listed below. For UL, a signalling chart is shown in Figure 5.1.1-1. For DL, a corresponding signalling chart is shown in Figure 5.1.1-2. In this example, UL synchronization is assumed.

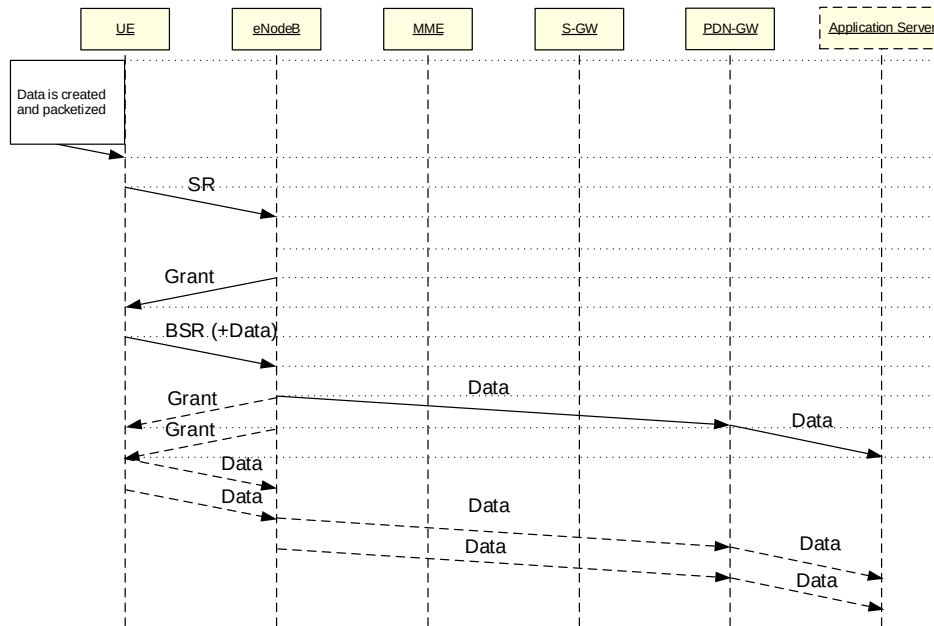


Figure 5.1.1-1. Overview of UL transmission delay. Does not include retransmissions.

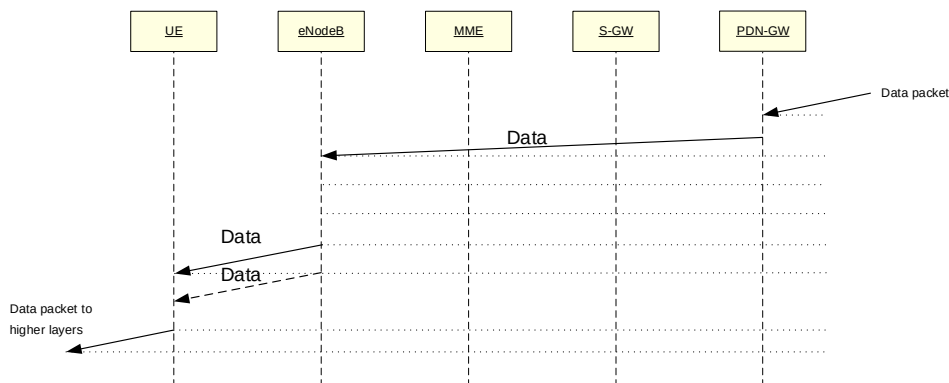


Figure 5.1.1-2. Overview of DL transmission delay. Does not include retransmissions

Grant acquisition

A UE with data to send must send a Scheduling Request (SR) and receive a scheduling grant before transmitting the data packet. In order to send a SR it must wait for a SR-valid PUCCH resource and a corresponding scheduling grant transmitted to the UE in response. When the grant is decoded the data transmission can start over PUSCH.

Random access

If the UL timing of a UE is not aligned, initial time alignment is acquired with the random access procedure. The time alignment can be maintained with timing advance commands from the eNB to the UE. However, it may be desirable to stop the maintenance of UL time alignment after a period of inactivity, thus the duration of the random access procedure may contribute to the overall latency in RRC_CONNECTED. The random access procedure also serves as an UL grant acquisition mechanism (Random Access based Scheduling Request). Therefore, for cases where random access is needed, no separate PUCCH based SR procedure/step is needed.

Transmission Time Interval (TTI)

The transmission of a request, grant, or data is done in subframe chunks with a fixed duration (1 ms), which is the source of a delay per packet exchange between the UE and the eNodeB.

Processing

Data and control need to be processed (e.g., encoded and decoded) in the UE and eNodeB. Data processing is a source of processing delays, which are proportional to the TB size. The processing of control information is typically less dependent on TB size.

HARQ Round Trip Time (RTT)

For UL transmission in FDD, the hybrid-ARQ acknowledgement for a packet received by the eNodeB in subframe n is reported in subframe $n+4$. If a retransmission is needed by the UE this is done in subframe $n+8$. Thus the hybrid-ARQ RTT is 8 ms for FDD UL. For TDD, RTT depends on TDD Configuration.

The RTT for DL transmissions is not specified in detail, as the HARQ scheme is asynchronous. The HARQ feedback is available at subframe $n+4$ in FDD, and retransmissions can typically be scheduled in subframe $n+8$ or later if needed.

Core / Internet

In the core network, packets can be queued due to congestion and delayed due to transmission over backhaul links. Internet connections can be congested and therefore add to the experienced end-to-end packet delay. EPC and/or Internet delays vary widely. In the context of latency reductions, it is reasonable to assume that latency performance of the transport links is good.

Note: In the context of evaluations in this study, a Core/internet latency component between 1-20 ms is assumed.

5.1.2 Handover latency [11]

Figure 5.1.2-1 below shows intra E-UTRAN handover procedure (Figure 10.1.2.1.1-1 of [12]) with the focus on handover execution phase, i.e., only step 7 to step 11.

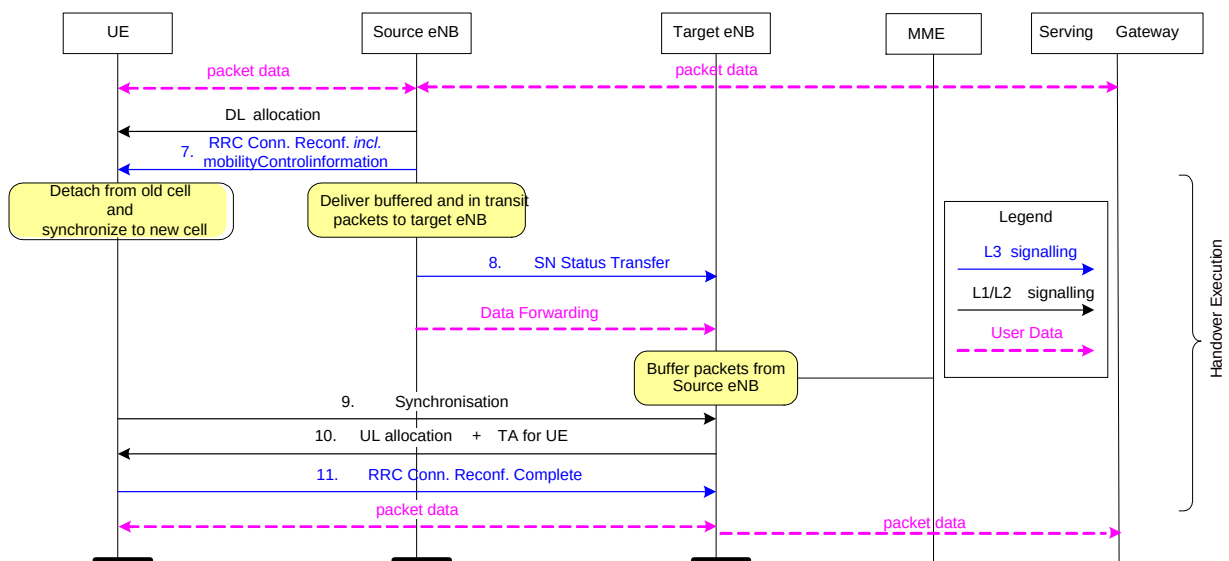


Figure 5.1.2-1: Handover execution procedure

Step 7: RRC Connection Reconfiguration Incl. *mobilityControlInfo*

In this step, the UE receives the *RRCCConnectionReconfiguration* message with necessary parameters (i.e. new C-RNTI, target eNB security algorithm identifiers, and optionally dedicated RACH preamble, target eNB SIBs, etc.) and is commanded by the source eNB to perform the HO [12]. RRC procedure delay includes RRC Connection Reconfiguration including *mobilityControlInfo* as well as related reconfigurations, including [12] [13]:

- Layer 2 reset / reconfiguration
 - Reset MAC.
 - Re-establish/reconfigure PDCP and RLC for all RBs that are established.

- Enable integrity protection and ciphering of RRC messages.
- Layer 3 reconfiguration (e.g. measurement configuration)

As per section 11.2 of TS 36.331 [13], for handover, the maximum allowed delay for RRC procedure is 15 ms.

Step 8: SN Status Transfer

In this step, the source eNB sends the SN STATUS TRANSFER message to the target eNB to convey the uplink PDCP SN receiver status and the downlink PDCP SN transmitter status of E-RABs for which PDCP status preservation applies (i.e. for RLC AM).

Since this is eNB to eNB signalling which is not related to air interface and it can be done in parallel with step 9 below, the delay contribution of this step can be considered negligible total handover latency.

Step 9: Synchronization

After receiving handover command (i.e., Step 7: *RRCConnectionReconfiguration* message including *mobilityControlInfo*), the UE may perform the following in this step [12] [13]:

- Physical layer synchronization and reconfiguration
 - Start synchronizing to the DL of the target PCell.
 - Reconfigure physical layer.
 - Access the target cell via RACH (following a contention-free procedure if a dedicated RACH preamble was indicated in the *mobilityControlInfo*, or following a contention-based procedure if no dedicated preamble was indicated)
- Layer 2 reconfiguration
 - Security key update: UE derives target eNB specific keys and configures the selected security algorithms to be used in the target cell.

According to TS 36.133 [14], when the UE receives a RRC message implying handover the UE shall be ready to start the transmission of the new uplink PRACH channel within D_{handover} seconds from the end of the last TTI containing the RRC command, where D_{handover} is the sum of RRC procedure delay and the “interruption time”. The interruption time is defined as the time between end of the last TTI containing the RRC command on the old PDSCH and the time the UE starts transmission of the new PRACH, excluding the RRC procedure delay.

Interruption time includes:

- Target cell search
- UE processing time for RF/baseband retuning, derive target eNB specific keys, configure security algorithm to be used in target cell
- RACH procedure related (uncertainty delay to acquire RACH opportunity followed by PRACH preamble transmission)

Step 9.1: Target cell search: As, in most cases, target cell is selected based on UE measurement reports and can be assumed to be “known”, the delay due to this step can be considered to be 0 ms for this study.

Step 9.2: UE processing time for RF/baseband re-tuning, security update: While the exact value can vary significantly based on various parameters, for the purpose of this SI, we consider UE processing time (for RF/baseband retuning, derive target eNB specific keys, configure security algorithm to be used in target cell) to be 20 ms as defined in TS 36.133.

Step 9.3: Delay to acquire first available PRACH in target eNB: Considering a typical RACH configuration where PRACH is available every 5 subframes, the minimum delay for this step 0.5ms and a typical delay would be 2.5 ms.

Step 9.4 PRACH preamble transmission: The last delay element in step 9 is 1 subframe required for PRACH preamble transmission.

Step 10: UL Allocation + TA for UE

In this step, the target eNB responds with UL allocation and timing advance. This corresponds to RAR from target eNB.

Assuming LTE FDD and that subframe number is continuously numbered, if UE sends RACH preamble in subframe n , following current specification, eNB can send RAR as early as in subframe $n+3$ (section 5.1.4 of TS 36.321[15]).

Assuming that the grant decoding and/or TA delay is not included in this step, the minimum delay of this step would be 3ms and typical/average delay would be 5ms.

Step 11: UE sends RRC Connection Reconfiguration Complete

When the UE has successfully accessed the target cell, the UE sends the *RRCConnectionReconfigurationComplete* message (C-RNTI) to confirm the handover, along with an uplink Buffer Status Report, whenever possible, to the target eNB to indicate that the handover procedure is completed for the UE. The target eNB verifies the C-RNTI sent in the *RRCConnectionReconfigurationComplete* message. The target eNB can now begin sending data to the UE [12].

According to section 6.1.1 of TS 36.213 [8], UE can then send *RRCConnectionReconfigurationComplete* as early as after $k_1 \geq 6$ subframes, i.e., the delay of this step is typically 6 ms. This includes UE Processing Delay (decoding of scheduling grant and timing alignment + L1 encoding of UL data) and transmission of RRC Connection Reconfiguration Complete.

5.2 Current performance

5.2.1 UL and DL latency

As an example, a simple assessment of important sources of latency for an UL transmission is presented in Table 1. Assuming Rel-8 functionality the average waiting time for a PUCCH at a periodicity of 10 ms is 5 ms, leading to a radio access latency sum of 17 ms in the example. With a SR period set to 1 ms, the average waiting time is reduced to 0.5 ms, which would lead to a sum of 12.5 ms in this example. Corresponding values for a DL transmission are given in Table 2.

Table 5.2.1-1. Typical radio access latency components (Rel. 8/Rel. 9) for an UL transmission from a UE without a valid uplink grant

Component	Description	Time (ms)
1	Average waiting time for PUCCH (10 ms SR period/1ms SR period)	5/0.5
2	UE sends Scheduling Request on PUCCH	1
3	eNB decodes Scheduling Request and generates the Scheduling Grant	3
4	Transmission of Scheduling Grant	1
5	UE Processing Delay (decoding of scheduling grant + L1 encoding of UL data)	3
6	Transmission of UL data	1
7	Data decoding in eNodeB	3
	Total delay [ms]	17/12.5

Table 5.2.1-2. Typical radio access latency components (Rel. 8/Rel. 9) for a DL transmission

Component	Description	Time (ms)
1	processes incoming data	3
2	TTI alignment	0.5
3	Transmission of DL data	1
4	Data decoding in UE	3
	Total delay [ms]	7.5

From the tables it can be seen that grant acquisition delay, transmission and data processing times are additive.

5.2.2 Handover latency [11]

As another example, based on discussion in Section 5.1.2, a simple assessment of sources of latency during handover execution is presented in Table 5.2.2-1.

Table 5.2.2-1. Minimum/Typical radio access latency components (Rel. 8/Rel. 9) during handover

Component/ Step	Description	Time (ms)
7	RRC Connection Reconfiguration Incl. <i>mobilityControlInfo</i>	15
8	SN Status Transfer	0
9.1	Target cell search	0
9.2	UE processing time for RF/baseband re-tuning, security update	20
9.3	Delay to acquire first available PRACH in target eNB	0.5/2.5
9.4	PRACH preamble transmission	1
10	UL Allocation + TA for UE	3/5
11	UE sends RRC Connection Reconfiguration Complete	6
	Minimum/Typical Total delay [ms]	45.5/49.5

It should be noted that above values assume successful transmission at first attempt. This may not always be true especially for handover scenarios where channel quality may be degraded. The actual delay values can be higher if some steps require retransmissions.

Based on above discussion, we see that total latency during HO process consists of various elements, as depicted in Figure 5.2.2-1. Service interruption time in handover can be defined as the duration between the time when UE stops transmission/reception with the source eNB and the time when target eNB resumes transmission/reception with the UE.

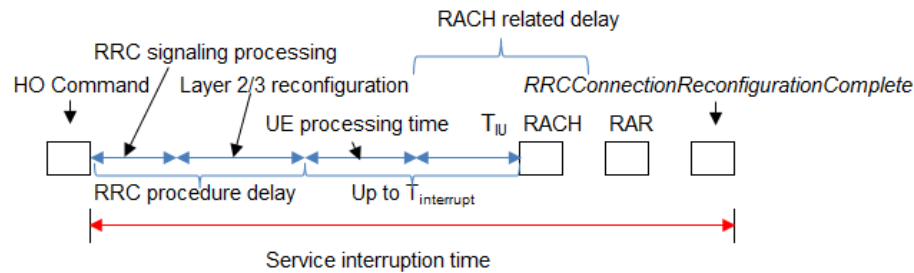


Figure 5.2.2-1: Service interruption time in handover

Broadly, the various delay components fall into one of the following three categories:

- RRC procedure delay, including RRC signalling processing (step 7)
- UE processing time, including delay for RF/baseband retuning, derive target eNB specific keys, configure security algorithm to be used in target cell (step 9.2)
- RACH procedure and *RRCConnectionReconfigurationComplete*, including delay to acquire first RACH occasion in the target cell (steps 9.3 to 11)

Among these, the following steps contribute to a major portion of total delay and can be addressed for possible latency reduction:

- RACH procedure including delay to acquire first available PRACH in target cell, PRACH preamble transmission and UL allocation + TA (steps 9.3, 9.4, 10),
- UE processing time after RA procedure including decoding of scheduling grant and timing alignment + L1 encoding of UL data, and transmission of *RRCConnectionReconfigurationComplete* (step 11).

5.3 Existing means to limit latency

With a short SR period, e.g. 1ms, the control plane overhead is increased which may reduce resource efficiency as more PUCCH resources in the cell to support the same number of users is needed. In addition, PUCCH resources are assigned and reconfigured with dedicated RRC signaling.

Pre-scheduling of scheduling grants uses PDCCH resources, and the granted PUSCH resources cannot be used by other UEs, which may limit the radio resource utilization. Further, the UE is expected to send a zero padded transmission also if the buffer of the scheduled UE is empty.

With SPS, periodic UL/DL resources can currently not be configured more frequently than every 10 subframes. Also with UL SPS, the UE is expected to send zero padded transmissions that may come with associated inefficient UE battery performance and increased UL interference.

6 Scenarios, Applications and Use Cases

A number of existing and new applications will benefit from reduced latency. Example of existing applications that would be positively impacted by reduced latency in terms of increased perceived quality of experience are TCP based applications, gaming, real-time applications like VoLTE/OTT VoIP and video telephony/conferencing.

For new applications it is expected that some will be increasingly delay critical; examples include remote control/driving of vehicles, augmented reality applications in e.g. smart glasses, or specific machine communications requiring low latency as well as critical communications.

It should be noted that applications/use cases have different characteristics and may be subject to different bottlenecks and benefit differently from particular improvements in latency.

7 Evaluation Structure and Assumptions

8 Solutions for latency reduction

8.1 Semi-Persistent Scheduling

With current Semi-Persistent Scheduling (SPS), the eNodeB may configure SPS periodicity via dedicated RRC signalling. Current minimum SPS periodicity is 10ms. Supporting a SPS periodicity of 1TTI is beneficial as this may reduce the latency of initial UL transmissions. This would allow UL transmission in consecutive subframes.

8.2 UL Grant reception

In current specifications, the UE sends a MAC PDU containing a MAC CE for padding BSR and optionally padding bits in response to an allocated UL dynamic or configured grant even if no data is available for transmission in the UE buffer and no other regular MAC CE is needed to be sent. It is beneficial to allow UEs to skip (most) dynamic and configured uplink grants if no data is available for transmission. With frequent UL grants, allowing skipping UL grants may decrease UL interference and improve UE battery efficiency. The UE will continue to send one or more regular MAC CE(s), if any. The eNB may enable skipping UL grants by RRC dedicated signalling.

8.2.1 Configured SPS activation and deactivation

The SPS resources are UE specific and may be reserved (configured and active) for a longer period in time. As such, it is in some cases considered useful for the eNB to be timely aligned with the UE state on activated or deactivated SPS resource(s), and if UL transmissions might occur in the granted SPS resource allocation instances.

An acknowledgement in response to a PDCCH grant indication activation could for example consist of an UE transmitted new MAC PDU containing zero MAC SDUs on the first SPS granted resource to indicate successful activation when the UE buffer is empty.

If the PDCCH grant indicating activation is not received, and UE has UL data to be transmitted, it is assumed that the UE would initiate a SR procedure if no other action is performed by the eNB.

For Deactivation, the eNB can indicate release of SPS resources on PDCCH. Also in these cases the eNB would in some cases benefit from being timely aligned with the UE state on activate or deactivated SPS resource(s).

Similarly to activation, a single transmission could be used also as deactivation acknowledgement as in instances of skipped transmissions as the absence of a UL transmission cannot be used as an implicit deactivation/release acknowledgement mechanism. Other means may thus be beneficial to specify.

Acknowledgment based Pro's

- 1 With an acknowledgment on the PDCCH grant activating or deactivating the SPS allocation, the eNB may be able to distinguish from cases where the PDCCH is missed by the UE.
- 2 With a higher degree of robustness of the activation/deactivation mechanism, the eNB may not have to repeat unnecessarily the SPS grant.
- 3 A activation and deactivation acknowledgement mechanism may allow for less conservative use of preconfigured SPS resources and may allow decreased PDCCH load as lower aggregation level can be used.
- 4 An acknowledged activation/deactivation of the SPS grant resources may reduce DRX misalignment.

Acknowledgement based Con's

- 1 Unnecessary UE battery consumption in instances of empty UE buffer at activation/deactivation;
- 2 DRX impacted by the additional UL transmission including the HARQ wait time (handling of inactivity timer /short cycle timer)

- 3 Increased uplink interference in instances where the SPS allocation is changed due to e.g. PDCCH activation/deactivation and where the UE buffer is empty.
- 4 New UE behaviour needed for deactivation/release acknowledgment.
- 5 DRX misalignment probability may be increased due to loss of ACK compared to if ACK is received correctly
- 6 SPS resources may be unused for a duration as a result of a missing ACK after successful PDCCH reception.

Acknowledgement free Pro's

- 1 Relies on the legacy behaviour that the UE in the rare cases of PDCCH loss (e.g. 1%) and non-empty buffer may initiate a SR procedure (alt. RACH procedure). The eNB may then re-initialise SPS or allocate a dynamic grant for the UE to compensate the PDCCH loss.
- 2 No additional UL transmissions that may cause interference in instances where the SPS allocation is changed due to e.g. PDCCH activation/deactivation and the UE buffer is empty.
- 3 UE may go earlier to DRX
- 4 UE behaviour simplified and no new behaviour at deactivation is needed.

Acknowledgement free Con's

- 1 Increased PDCCH aggregation level may need to be used which increases the PDCCH resource usage
- 2 SPS resources may be unused until a SR procedure has been initiated by the UE missing the PDCCH
- 3 SPS deactivation on PDCCH may need repetition to have robust SPS resource release
- 4 SPS resources are not released after release indication and are only detected at first periodic grant allocation opportunity after the release signalling time instance.

8.3 Handover Latency

One possible area of improvement where several existing and new applications could benefit from is to minimize latency during handovers for a mobile UE. Although TCP based applications could recover from errors introduced by poor radio conditions during handovers, the delay introduced by handovers for real time applications is least desired. Two potential solutions to reduce latency during HO are provided below.

8.3.1 Solution 1: RACH-less handover

The solution of RACH-less handover can be introduced when the source cell, the target cell and the UE are synchronized. In a synchronized network, it is assumed that subframe boundary between the source cell and target cell are aligned. One option is that at a mutually agreed time (e.g. SFN), the UE switches from source cell to target cell, without requiring random access procedure. Another option is that the UE follows the legacy handover procedure but skips the RACH related steps. A RACH attempt procedure during handovers typically takes ~10~12 ms. An average handover procedure takes ~40~50 ms to complete. Eliminating ~10~12 ms of RACH delay during a handover procedure can significantly reduce the data interruption during handovers and improve the user experience.

The synchronization may be achieved between the two eNB cells over X2 signalling and the UE via RRC signalling. Based on the fact that all three nodes are in sync, the source cell stops DL transmission to the UE, the target cell provides an uplink grant to the UE, and the UE acquires the target.

Target cell TA without RACH

One of the main purposes of RACH procedure during HO is to obtain target cell TA. In absence of RACH procedure, the UE may be able to obtain the target cell TA without explicit TA command when the source cell and the target cell are time synchronized. As illustrated in Figure 8.X.2-1, the UE first obtains the DL propagation delay difference between the source cell and the target cell (i.e., $T_1 - T_2$). Assuming the UL propagation delay is the same as the DL propagation delay, the UE can derive the target cell TA from the source cell TA by:

$$TA_{\text{target}} = TA_{\text{source}} - 2(T_1 - T_2).$$

NOTE: The accuracy of the target cell TA obtained by this method needs to be further checked by RAN4.

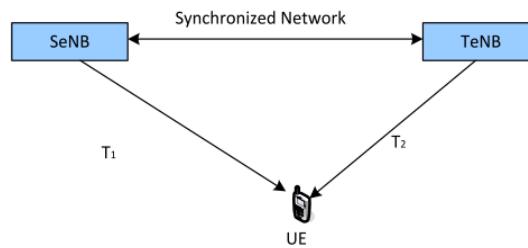


Figure 8.3.1-1: Obtain target cell TA

UL grant allocation for handover command response

Another purpose of RACH procedure during HO is to obtain UL grant for the transmission of the handover command response (i.e., RRC Connection Reconfiguration Complete message). In absence of RACH procedure in target eNB, allocation of UL grant is needed in the target cell. One option is UL grant pre-allocation in handover command. The pre-allocated UL grant can be kept valid within a period of time, starting from the time when the UE achieves synchronization with the target cell. Another option is UL grant allocation by dynamic scheduling in the target cell. The target cell allocates UL grant to the UE by dynamic scheduling from the time when it expects the UE to be available for scheduling (e.g., based on mutually agreed time, or sometime later after the handover preparation procedure which is subject to eNB implementation).

The initial value of PUSCH transmission power control is based on PRACH preamble power and total power ramp. If PRACH procedure is removed, power control in PUSCH should be modified. The impact in uplink power control needs to be studied by RAN1.

8.3.2 Solution 2: Maintaining Source eNB Connection during Handover

This solution can reduce data interruption during handover (HO) by not releasing the connection to the source eNB until handover is completed at the target eNB. Continuous transmission of user data from source cell well after handover procedure start can significantly improve the user experience.

In current LTE HO, the UE resets MAC and reestablishes PDCP upon receiving the HO command and thus communication with the source eNB is stopped. The data disruption will happen until the UE receives the first packet from the target eNB. The proposal is to delay the MAC and PDCP reset until the UE performs successful RACH at the target eNB. This requires that the UE monitor both source and target links simultaneously, which is similar to Dual Connectivity on the same frequency. To continue the data transmission with the source eNB, the UE should continue sending CSI and HARQ feedback to the source eNB. However, if the radio link quality becomes bad, this shouldn't force the UE to declare RLF since that would trigger RRC Connection Re-establishment. Therefore, RLM for the source eNB needs to be suspended until the end of handover (success or failure). The impact on uplink transmission should be evaluated if the UE still sends uplink signal in bad radio link quality due to suspension of RLM.

This solution can provide the additional benefit of returning back to the source eNB if handover fails. When RACH is completed successfully at the target eNB, the source eNB should be made aware that it can stop transmissions to the UE. This indication can come from the UE or from the target eNB with the X2 option being more efficient due to better reliability.

Under current specifications, the UE cannot transmit simultaneously to source cell and target cell on the same frequency although this may be feasible by using multiple RF chains at least on downlink while uplink may require changes such as TDM operation. Since this can impact physical layer procedures, this solution should also be studied by RAN1/RAN4.

8.4 Contention based PUSCH transmission

In the pre-scheduling scheme allowed by current specifications, the eNB will assign one separate UL grant for each UE in each pre-scheduling interval, and the assigned UL grant will be wasted if one UE has no available data to transmit during one pre-scheduling interval. For contention based PUSCH transmission, multiple UEs may share the same PUSCH resource (either dynamically granted or configured). Collision will happen if two or more UEs that share the same PUSCH resource perform the PUSCH transmission at the same time, and in this case the eNB may not be able to successfully decode all of the PUSCH transmissions. Contention based PUSCH transmission allows more efficient

PUSCH resource utilization compared to the existing pre-scheduling scheme. However, as a result of collision, the potential retransmissions can result in increased latency for colliding UEs.

8.4.1 Solution 1 [16]

This contention based PUSCH transmission scheme is featured with UE identification even if collision happens. If the UE has available data to transmit and no collision happens, the initial transmission is identical with the existing PUSCH transmission. Nevertheless, when the initial transmission fails, the eNB needs to schedule the retransmission on a dedicated PUSCH resource to avoid the potential collision. If the UE has available data to transmit and unfortunately collision happens, for each of the colliding UE, since the eNB can distinguish it, the eNB can either first respond by PHICH ACK to hold the uplink transmission and later schedule the retransmission on a dedicated PUSCH resource, or immediately schedule the retransmission on a dedicated PUSCH resource, so as to avoid further collision. A UE configured with contention based PUSCH transmission also needs to be configured with UL grants skipping if no available data to transmit (as described in section 8.2) to minimize the potential collision.

An example of distinguishing the colliding UEs is DMRS-based UE identification. The eNB may allocate different DMRS resources (i.e. different DMRS Cyclic Shifts) to different UEs that share the same PUSCH resource, so that the eNB can identify the exact UE that performed the PUSCH transmission even if collision happens. In the current specification, there are 8 different DMRS resources. In some scenarios, the eNB may only use part of them (e.g. 4 out of 8) to maximize the orthogonality.

If the contention based PUSCH resource is assigned by dynamic scheduling, then there is no impact to the current specifications, as the DMRS Cyclic Shift can be provided by the DCI for uplink scheduling. If the contention based PUSCH resource is assigned during SPS activation, considering that the DMRS Cyclic Shift provided in the DCI for SPS activation is fixed to '000', RRC specification needs to be updated so that the eNB can provide the DMRS Cyclic Shift to the UE in SPS configuration.

8.4.2 Solution 2 [17]

The characteristics of this CB PUSCH mechanism are given as below:

- More than one UE shares the same CB resource indicated by CB grant;
- If UE needs to transmit uplink data, it monitors CB grant, and transmits the data on the indicated CB resource; otherwise, UE is not required to monitor CB grant;
- If only one UE transmits data on the CB resource, there is no collision:
 - Based on the received data eNB can acquire the related UE information, and gives the response to that UE;
- Otherwise, if more than one CB UE transmits data simultaneously, the collision happens:
 - For eNB, it cannot successfully decode the data received on CB resource, and does not give any feedback to UE;
 - For UE, it can realize the collision by not receiving any feedback, and perform the following actions:
 - Flush the HARQ buffer for this CB transmission;
 - Indicate the transmission failure information to RLC, and RLC performs the related RLC PDU retransmission;
 - Perform backoff mechanism, and the next CB transmission can be performed after a random backoff time;

Note: Here we assume the value of backoff time is between 0 and 10ms.

Note: HARQ retransmission is not supported by CB-PUSCH transmission.

8.5 Reduced TTI and processing time

9 Performance Evaluation

Editors Note: The study evaluation includes resource efficiency, including air interface capacity, battery lifetime, control channel resources, specification impact and technical feasibility. Both FDD and TDD duplex modes are considered.

9.1 Protocol evaluations on TTI reduction and Fast UL

TCP Slow Start

TCP Slow-start is in general part of the congestion control mechanism used by TCP and is used in conjunction with other algorithms to avoid network congestion. Slow-start begins initially with a congestion window size, and it will be increased with each acknowledgment received, effectively increasing the window size each round trip time (RTT). The transmission rate will be increased with slow-start algorithm until either a loss is detected, or the receiver's advertised window is the limiting factor, or the slow-start threshold is reached. If a loss event occurs, TCP assumes that it is due to network congestion and takes steps to reduce the offered load on the network [4].

9.1.1 Simulation 1

TCP slow-start behavior for FTP file download based on reduced TTI and reduced SR periodicity [4]

In this simulation it is assumed that the core network and the Internet delay is very small and is independent of the TTI values. The assumed latency between eNB and TCP Server is 1-2 ms.

The average DL data-rate the same for different scenarios. This is achieved by scaling the TBS sizes according to the duration of TTI compared to the current 1ms TTI.

For the FTP based file download using TCP protocol, TCP New Reno is used. File size is kept so as to achieve the saturation in data-rate so that the conditions when UE's perceived instantaneous as well as overall data-rate reaches at least 90% of the saturation rate can be determined.

The instantaneous throughput is calculated as 25ms moving window average of TCP layer throughput. Aggregate user perceived throughput (UPT) is calculated as total TCP layer data downloaded divided by the time elapsed. Initial time for setting up the network and for the UE to attach to network is removed as offset from the results such that the FTP request is sent to the server from the UE at time 0.

A saturation data-rate of approximately 21Mbps (example scenario is 3 CA 20MHz each, cell spectral efficiency ~3.5bps/Hz, 10 UEs/cell) is assumed. MCS index of UE is set such that on an average this data-rate is achieved when TCP has fully ramped up (channel adaptation is disabled).

The results for TTI reduction in this simulation do not take into account possible overhead due to e.g. additional L1/2 overhead.

Results simulation 1

Figure 3 shows the comparison of 1ms TTI with 0.5ms and 0.1ms TTI scenarios. The SR periodicity is 5 TTI in each case. As expected the TCP ramps up quickly with smaller TTI durations compared to 1ms TTI.

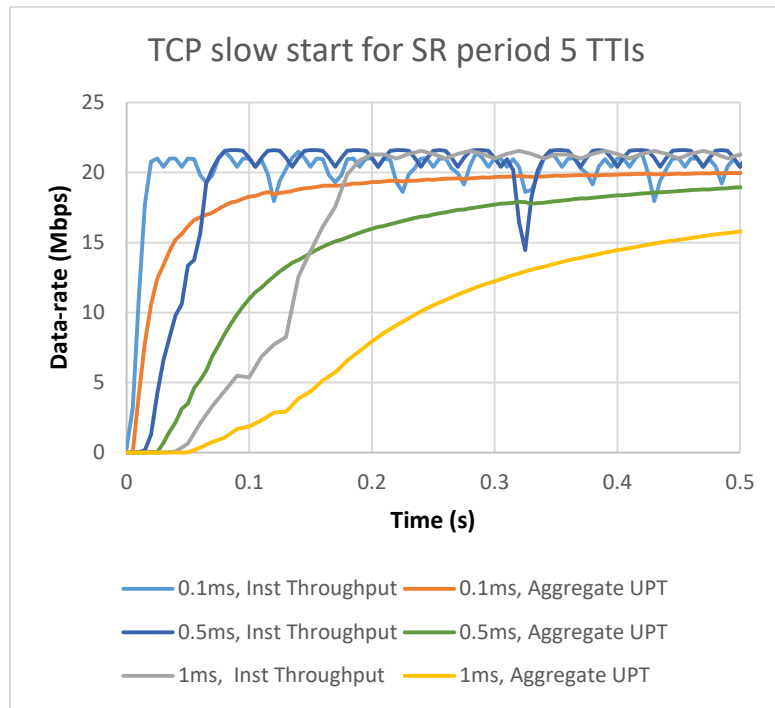


Figure 3 Effect of reduced TTI on TCP slow-start behaviour for SR periodicity of 5 TTIs

Figure 4 shows the comparison of TCP ramp-up behaviour with SR periodicity of 1 TTI for different TTI values.

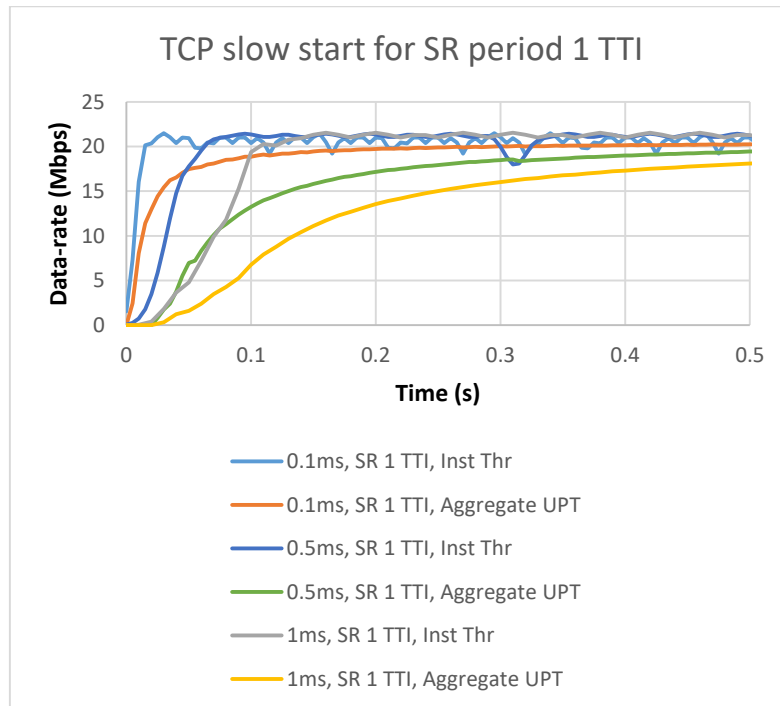


Figure 4 Effect of reduced TTI on TCP slow-start behaviour for SR periodicity of 1 TTI

Table 3 shows the comparison of the six scenarios described above. It is observed that, the slow-start ramp-up time to 90% saturation throughput reduces by 44% when SR periodicity of 1 TTI is used for the 1ms TTI duration compared to SR periodicity of 5 TTIs. The ramp-up time is observed to reduce by 92% when TTI is reduced from 1ms to 0.1ms keeping the same SR periodicity (5 TTIs). Note that because of reduction in TTI duration, the periodicity of SR opportunities implicitly increases by the same factor when periodicity is fixed in terms of number of TTIs. Therefore, to

compare the TCP performance with the same absolute periodicity of SR opportunities among the considered scenarios, we can compare 0.5ms TTI with SR periodicity 1 TTI and 0.1ms TTI with SR periodicity 5 TTI, where we observe 73% reduction in TCP slow-start time and 63% reduction in minimum file size for TCP ramp-up to 90% of the saturation rate as a result of TTI reduction from 0.5ms to 0.1ms.

Observation 1. When core network and the Internet latencies are minimal, air-interface latency becomes the dominant factor.

Observation 2. TCP slow-start ramp-up time reduces significantly with smaller TTI durations compared to 1ms TTI.

Observation 3. TCP slow-start ramp-up time reduces with reduced SR periodicity.

Table 3 Comparison of TCP slow-start performance

TTI, SR periodicity in TTIs	Time for TCP ramp-up to 90% of saturation rate*	Min file size for TCP ramp-up to 90% of saturation rate	Time taken for UPT to be 90% of saturation rate*	Min file size for UPT to be 90% of saturation rate
1ms, 5 TTI	180ms	156KB	1.31s	3.12MB
0.5ms, 5 TTI	70ms	64KB	485ms	1.56MB
0.1ms, 5 TTI	15ms	20KB	150ms	366KB
1ms, 1 TTI	100ms	93KB	720ms	1.72MB
0.5ms, 1 TTI	55ms	54KB	380ms	908KB
0.1ms, 1 TTI	10ms	15KB	100ms	248KB

*Note: minimum time granularity for TCP throughput observations in Table 3 is 5ms.

Table 3 also shows that the minimum file size for the users to realize aggregate UPT of 90% of saturation rate decreases by about the same factor (45%) when SR periodicity of 1 TTI is used for 1ms TTI duration compared to SR periodicity of 5 TTIs. The same performance metric (minimum file size for UPT to be 90% of saturation rate) is observed to reduce by 88% when TTI is reduced from 1ms to 0.1ms keeping the same SR periodicity (5 TTIs). Compared to 1ms TTI with SR periodicity of 5 TTIs, 0.1ms TTI with SR periodicity of 1 TTIs results in 13x reduction in minimum file size required to realize UPT of 90% of saturation rate.

Table 4 shows the average air-interface latency for UEs in connected mode. Average downlink latency and latency for the UE initiated uplink data transmission for a UE with uplink synchronization are calculated following the timing analysis of [3]. The details of the calculation are shown in Appendix A. For this calculation, we assume that the processing times at the eNB and UE can be reduced by the same factor as the TTI reduction; and two cases for UL and DL HARQ transmissions – a) no HARQ retransmissions are required (i.e., error-free), and b) 10% error probability for initial transmission and up to 1 HARQ retransmission. For the UL, we also separate the latency due to SR related steps and actual UL data transmission where we assume that the UE processing time is divided into half for scheduling grant decoding and L1 data encoding. To be consistent in comparison of DL and UL, we add the eNB processing delay in the UL after the UL data is received by the eNB.

Table 4: Average Air-interface latency

TTI, SR periodicity in TTIs	Average air-interface latency							
	DL		UL					
	error free	10% error	error free			10% error		
			SR- related	Data Tx	Total	SR- related	Data Tx	Total
1ms, 5 TTI	4ms	4.8ms	9ms	4ms	13ms	9ms	4.8ms	13.8ms
0.5ms, 5 TTI	2ms	2.4ms	4.5ms	2ms	6.5ms	4.5ms	2.4ms	6.9ms
0.1ms, 5 TTI	0.4ms	0.48ms	0.9ms	0.4ms	1.3ms	0.9ms	0.48ms	1.38ms
1ms, 1 TTI	4ms	4.8ms	7.0ms	4ms	11ms	7.0ms	4.8ms	11.8ms
0.5ms, 1 TTI	2ms	2.4ms	3.5ms	2ms	5.5ms	3.5ms	2.4ms	5.9ms
0.1ms, 1 TTI	0.4ms	0.48ms	0.7ms	0.4ms	1.1ms	0.7ms	0.48ms	1.18ms

It is observed from the latency values in Table 4 and TCP ramp-up performance results in Table 3 that the TCP ramp-up performance is highly correlated, as expected, to the average air-interface latency, i.e., the ramp-up is faster if the air-interface latency is lower. In addition, the same benefits in terms of TCP slow-start performance can be observed with latency reduction in general regardless of how the lower latency is achieved.

Observation 4. Reduced air-interface latency can help to achieve higher user perceived throughput even for smaller TCP sessions.

Observation 5. TCP slow-start performance benefits are highly correlated with the UL air-interface latency but independent of the method how the faster UL access is achieved.

It is also observed that SR-related latency in the UL is the dominant factor in average latency values. It is noted that SR-related delay includes delay due to SR periodicity; uplink scheduling delay and uplink grant transmission. Therefore, it is clear that solutions, which reduce or completely remove the SR-related delay, will improve the TCP slow-start behaviour significantly.

9.1.2 Simulation 2

Capacity and throughput gain with 0.5ms TTI [5]

Table 5 lists the distribution (i.e. probability) of the number of PRB(s) occupation for different packet size. This PRB distribution is a result from system simulations, using 0.5ms TTI length. In Table 2.2-1, 1 PRB means one PRB for 1ms TTI length (i.e. existing PRB concept), and 0.5 PRB means half of the existing PRB. Taken the packet size of 32 Bytes as one example, the probability that it occupies 0.5 PRB is 25.02%.

Table 5: Distribution of # PRB for different packet size

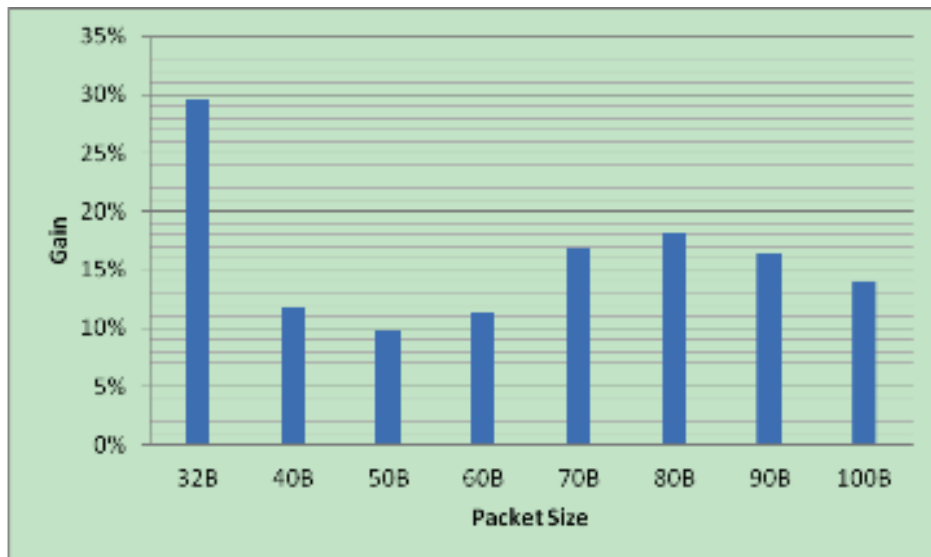
Index	i =1	i=2	i=3	i=4	i=5	i=6	i=7	i=8	i=9
	0.5 PRB	1 PRB	1.5 PRB	2 PRB	2.5 PRB	3 PRB	3.5 PRB	4 PRB	≥4.5 PRB
32 Bytes	25.02%	55.10%	12.65%	3.44%	1.71%	1.21%	0.14%	0.25%	0.21%
40 Bytes	5.48%	64.48%	16.33%	7.17%	3.04%	1.38%	1.01%	0.17%	0.06%
50 Bytes	2.03%	56.94%	20.53%	10.62%	3.68%	2.70%	0.96%	0.94%	0.64%
60 Bytes	1.29%	43.15%	25.80%	13.84%	5.64%	3.35%	2.48%	1.66%	0.41%
70 Bytes	3.68%	28.37%	33.06%	14.23%	7.81%	3.33%	3.09%	2.28%	1.73%
80 Bytes	0.39%	8.79%	45.29%	16.56%	10.57%	5.38%	3.19%	3.81%	1.29%
90 Bytes	0.34%	6.08%	37.82%	21.29%	12.21%	5.92%	4.89%	2.14%	2.81%
100 Bytes	0.25%	3.88%	31.82%	22.68%	11.68%	11.36%	4.25%	3.17%	2.29%

With 0.5ms TTI, different packet size will obtain different gains, and the gain for different packet size could be calculated as below:

$$TotalGain = \sum prob_i * Gain_i$$

For example, for the packet size of 32 Bytes, if one packet occupies 0.5 PRB with 0.5ms TTI, it means it will occupy 1 PRB with 1ms TTI. In this case, the capacity $Gain_1$ is $(1-0.5)/0.5 = 100\%$. If one packet occupies 1 PRB with 0.5ms, it means it will also occupy 1 PRB with 1ms TTI. In this case, the $Gain_2$ is 0%. Similarly, the $Gain_3$ is $(2-1.5)/1.5 = 33.33\%$.

Figure 5 illustrates the system capacity gain for different packet size provided by 0.5ms TTI. It could be observed that the packet size of 32Bytes can achieve the highest capacity gain, which is about 29.5%.

**Figure 5: Capacity gain for different packet size**

Observation 6. With 0.5ms TTI, system capacity gain can be achieved for small packet. For the packet size of 32Bytes, 29.5% system capacity gain can be achieved.

FTP throughput

Today, TCP is the dominating transport layer protocol used for varies applications over the Internet. For a FTP file with relative small size (e.g. 0.5Mbytes), the TCP slow start period is a significant part of the total transport period. During TCP slow start, the performance is latency limited. Shortened TTI over the air interface can accelerate the TCP ACK

from the peer entity to shorten the TCP slow start phase, so that the TCP throughput can be improved. Further more, shortened TTI over the air interface can also shorten the TCP RTT, which will also contribute to the increased TCP throughput.

Figure 6 illustrates the UPT (User Perceived Throughput) gain provided by 0.5ms TTI for FTP download, wrt different number of UEs per cell. The FTP file size is 0.5Mbytes. The simulation assumptions on radio aspects are provide in Annex A.2, the FTP traffic model is provided in Annex A.4, and the TCP parameters are provided in Annex A.3.

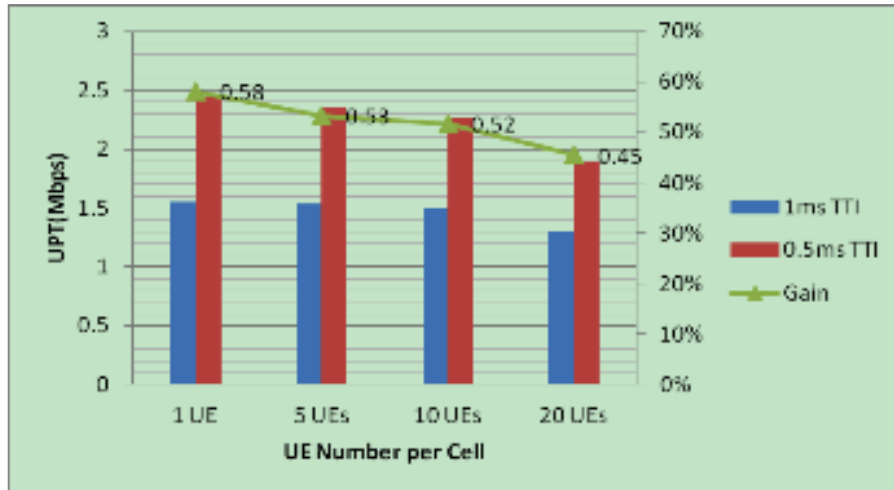


Figure 6: UPT gain for FTP download

It could be observed that 58% UPT gain can be achieved with 0.5ms TTI, if there is only one UE per cell. For increased number of UEs per cell, the UPT gain will be slightly decreased, as radio interface becomes the bottleneck.

Figure 7 shows how the TCP slow start phase can be promoted with 0.5ms TTI, taken one UE with moderate geometry as the example. It could be observed that the TCP window for the TCP transmission with 0.5ms TTI grows more quickly and than that with 1ms TTI in the TCP slow start phase.

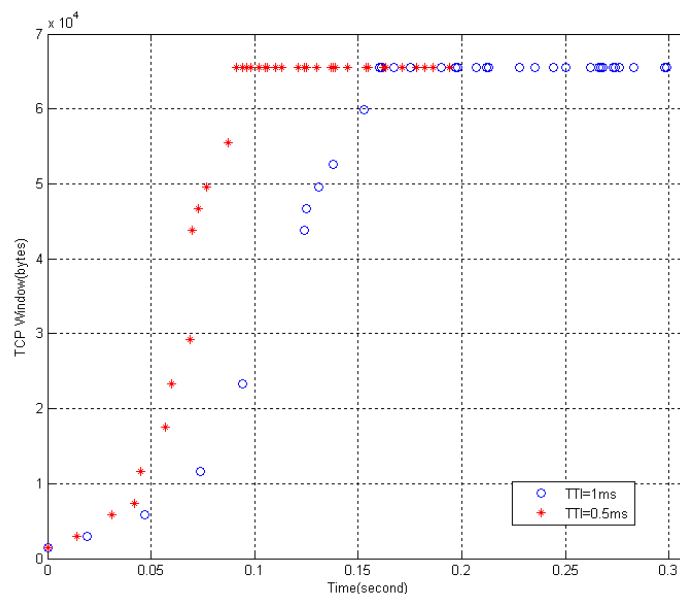


Figure 7: Accelerated TCP window by 0.5ms TTI

Observation 7. With 0.5ms TTI, FTP UPT gain can be achieved. For the FTP file size of 0.5Mbytes, 58% UPT gain can be achieved.

9.1.3 Simulation 3

Throughput and packet download time with reduced latency in LTE [3]

Figure 8 shows the potential gains in terms of increased download throughput and reduction of download time for reduced latency in LTE. In the simulation download file sizes (e.g. of an FTP item, downloaded with TCP) between 100 KB and 5MB were analysed.

For the LTE baseline a peak-latency performance when no retransmissions are needed is assumed (e.g. 0% HARQ BLEP). Thus, for the DL a delay of 4ms for transmission and decoding is assumed, and for the UL a 10ms PUCCH cycle is assumed, and a SR to grant period of 8ms, as well as 4ms for transmission and decoding, which leads to an average total of 17ms latency in the UL. For transport, core network and Internet a total delay of 10ms is assumed. Only the %-reduction of the radio delay is analysed in the figures. The download throughput is assumed to be 150Mbit/s and the uplink throughput 50Mbit/s.

One can see that e.g. for a radio latency reduction of 50%, the download time for a file of less than 1MB can be reduced by 25%. This is for an unchanged throughput of the LTE system.

It can further be noted that for the LTE baseline a HARQ block error probability of 0% was assumed. For higher block error probabilities, higher average LTE latencies need to be considered due to possible HARQ retransmissions. In this case, overall latency reductions could also be achieved e.g. by means of HARQ RTT reductions.

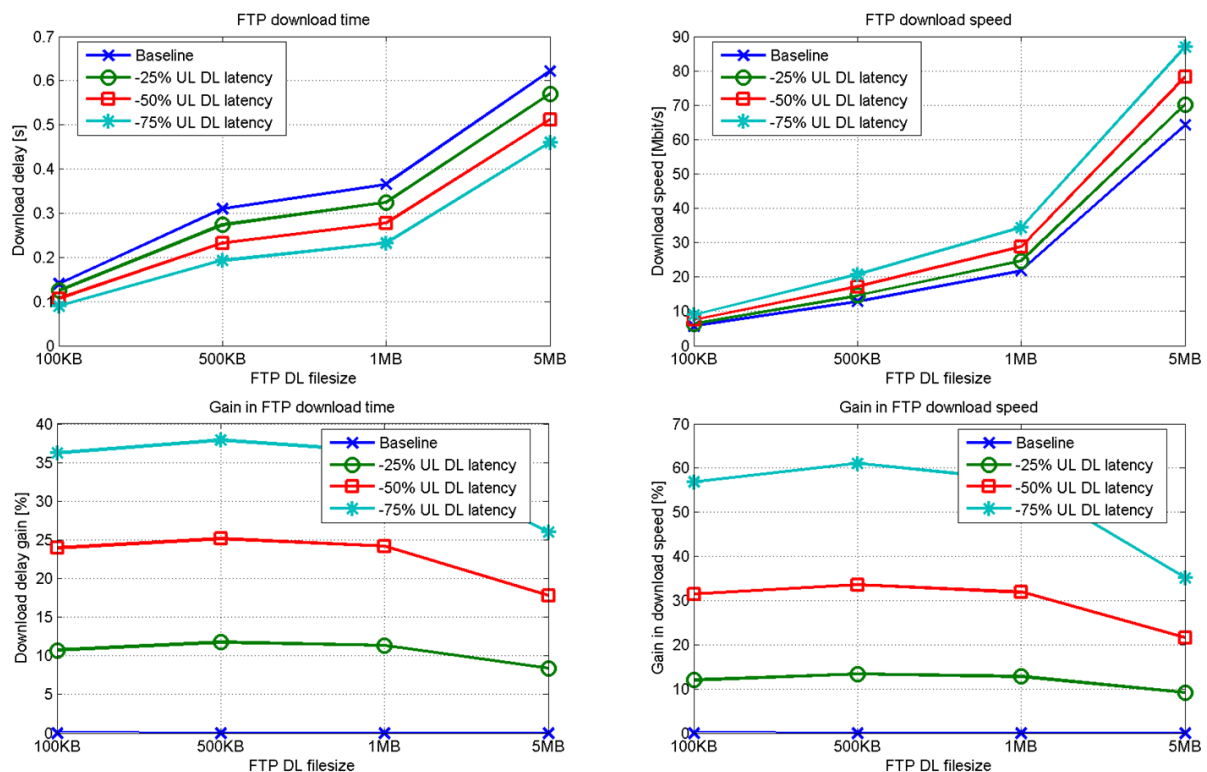


Figure 8. Simulated download delay, speed and relative gain in speed and delay (y-axes) for different file sizes (x-axis) and percentages of UL and DL latency reduction techniques

In the above, the maximum achievable LTE link throughput is assumed to be 150 Mbit/s in this simulation and the transport and Internet delay is 10 ms.

Observation 8. Considering the additive effect of grant acquisition, transmission and processing delay reductions, overall latency reduction and performance improvement can be significant.

9.1.4 Simulation 4

Latency evaluation results for TTI reduction and Fast UL [6]

9.1.4.1 Simulation assumptions

The performance of a single FTP download over TCP for file sizes between 100 kB and 5 MB is evaluated; in addition to a radio network delay a fixed delay of 10 ms from core, transport and Internet is assumed.

For the LTE radio part a system with a 10 TTI PUCCH cycle, an 8 TTI SR to grant period, and an 8 TTI HARQ round trip time (allowing for 8 HARQ processes) is assumed. For the LTE baseline, a TTI length of 1ms (14 OFDM/SC-FDMA symbols) is assumed.

With the technique of TTI shortening the TTI length is scaled to 7, 2, and 1 symbol. The TTI scaling in this evaluation also directly scales the HARQ RTT, so that a proportional scaling of the transport block size is achieved. For simplicity, it is further assume that a reduction of the TTI length also scales the SR to grant period, as well as the PUCCH cycle. For the shorter TTIs it is assumed that an additional DL and uplink overhead due to increased signalling applies, see Table 9.1.4-1. For the SR period one standard case (10 TTI) and one short case (1 TTI) is studied for comparison.

With a faster first UL grant, the UE is assumed to not perform a legacy PUCCH cycle nor SR to grant period, i.e. the grant acquisition delay is assumed to be 0ms.

In Table 9.1.4-1 below we summarize the simulation parameter setup used in this evaluation.

Table 9.1.4-1. TCP download simulation setup	
Simulation scenario	Static UE in a cell
DL and UL throughput	DL 150 Mb/s, UL 50 Mb/s
Number of OFDM/SC-FDMA symbols per TTI, same in both DL and UL	1, 2, 7, 14 corresponding to - HARQ RTT: of 0.57ms, 1.14 ms, 4 ms, 8 ms - SR to grant period: 0.57ms, 1.14, 4 ms, 8 ms - PUCCH cycle: 0.7ms, 1.4 ms, 5 ms, 10 ms
HARQ error probability	0.3 for 1 st transmission, 0.1 for 2 nd , 0.05 for 3 rd , 0.01 for 4 th , ...
Traffic model	FTP download over TCP (initial window: 10segments)
Core, transport and internet network delay	10ms
FTP file size	[100 kB, 500 kB, 1 MB, 5 MB]
SR period	[1 TTI, 10 TTI]
DL PHY overhead	7%, 7%, 25%, 25% for TTI length of 14, 7, 2, 1 symbols respectively
UL PHY overhead	14%, 14%, 33%, 50% for TTI length of 14, 7, 2, 1 symbols respectively

9.1.4.2 TTI shortening simulations

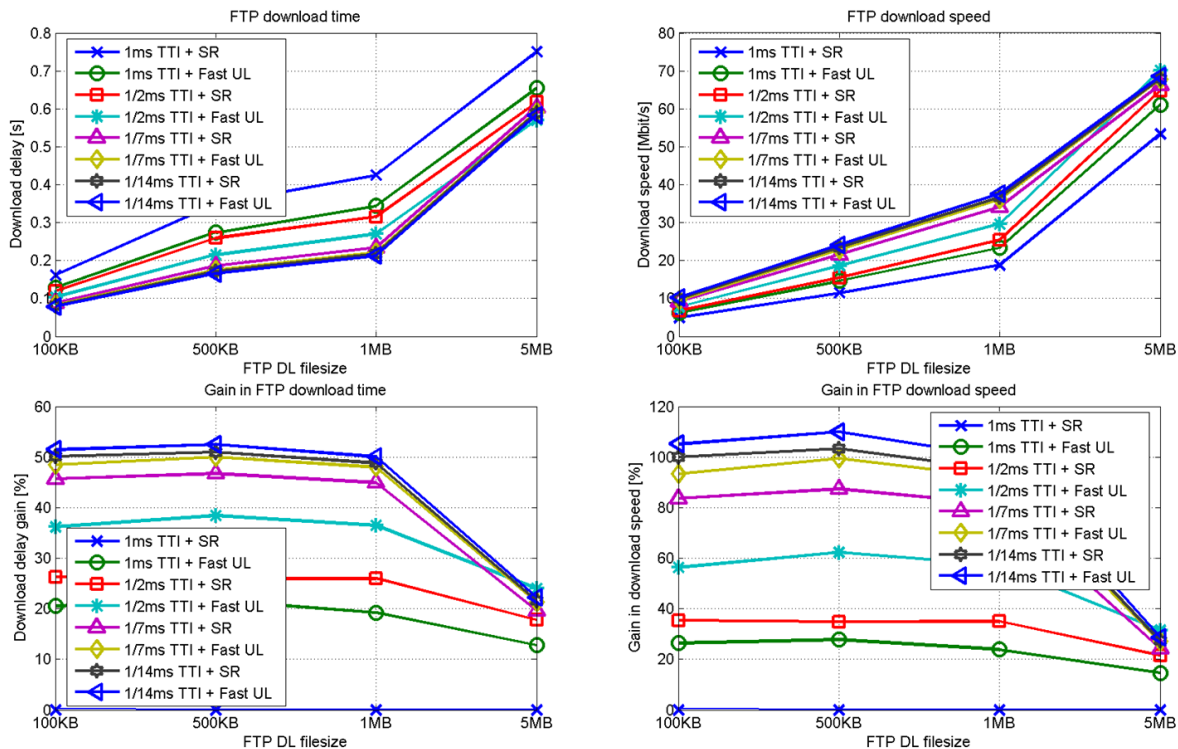


Figure 9.1.4-1: Download speed and delay, as well as relative gains in speed and delay compared to baseline for different file sizes and latency reduction techniques in LTE. The baseline is 1 ms TTI and SR. SR period is 10 TTI.

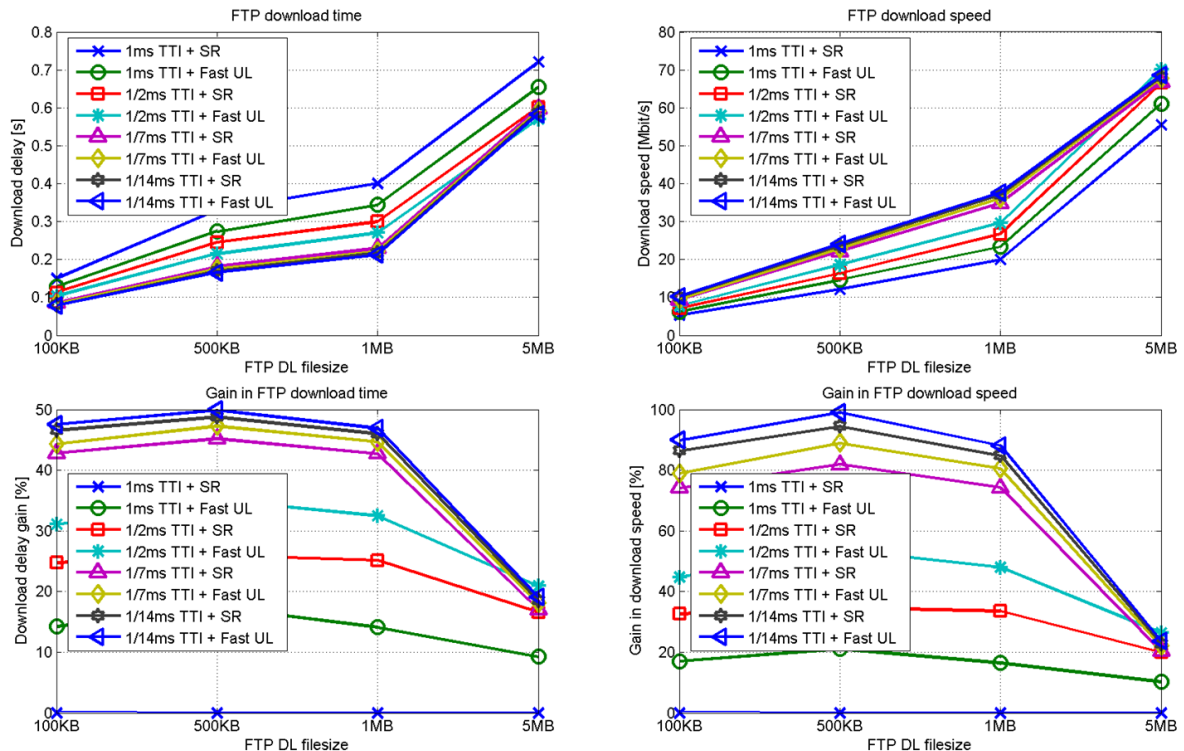


Figure 9.1.4-2: Download speed and delay, as well as relative gains in speed and delay compared to baseline for different file sizes and latency reduction techniques in LTE. The baseline is 1 ms TTI and SR. SR period is 1 TTI.

From the results in Figure 9.1.4-1, for a SR period of 10 TTI, it is clear that significant gains in download delay and FTP object bit rate can be achieved at smaller FTP file sizes with the two studied latency reduction techniques. Generally, the smaller the FTP file size is (for a given link level throughput), the larger the fraction of the TCP slow start phase on the overall download time. During the TCP slow start phase the maximum achievable throughput is not fully utilized, instead the performance depends on the RTT, i.e. directly on the latency.

The same trend can be seen in Figure 9.1.4-2, for a SR period of 1 TTI, but compared to the 10 TTI case the baseline is somewhat improved, and therefore the gains with shorter TTI and Fast UL are slightly reduced.

It becomes evident that a high gain in terms of overall download time reduction can be achieved with a faster first UL grant (green line); compared to the baseline the download time is reduced by 20% for smaller file sizes. A similar and somewhat higher reduction can be achieved with a 7 symbol TTI (0.5ms), i.e. 25% reduction (red line). It is noteworthy that these techniques can complement each other. With a combined use, reductions of about 35% can be achieved for the download time (cyan line).

Further improvements are achievable with further reduction of the TTI length. In the case of 2 symbol TTI (1/7ms) and with the use of fast UL grant as well (orange line), the total download time reduces by 50% at the smaller file sizes. For a 0.5 MB FTP download this corresponds to a reduction from a 0.35 s to 0.18 s download time. Reducing even further to 1 symbol TTI (1/14ms) provides only a minor additional benefit compared to the 2 symbol TTI. This is due to the assumed comparably large transport/internet/core delay of 10ms compared to the reducible RAN delay.

For file sizes below 5MB an improvement in performance can be achieved from shortening the TTI length further than 1/2ms, while at 5MB file size a TTI length of 1/2ms gives the best performance, due to the impact of higher DL overhead for TTI reductions beyond 1/2ms.

Observation 9. Reducing TTI length down to 1 or 2 symbols gives largest latency reduction for file sizes up to 1MB, and for larger file sizes a TTI length of 7 gives best latency performance.

Observation 10. Adding Fast UL gives an improved performance with both legacy and short TTI length in all studied cases.

Observation 11. The gain in both download speed and time is larger for smaller file sizes with all TTI lengths compared to the 5MB file size. Further, for the small file sizes, these gains are largest with 1 symbol TTI. These observations are due to the proportional length of TCP slow start compared to the total file download time.

9.1.4.3 Fast UL system simulations

The Fast UL grant mechanism was simulated in a system level simulator tool with FTP traffic and its performance was compared to that of SR grants and SPS grants. The simulator parameters are given in Table 9.1.4-2. Figure 9.1.4-3 shows the user FTP throughput for the different grant mechanisms. As can be seen in the figure, a gain of 30-40% can be achieved with fast UL grants compared to SR grants with a period of 10 TTI, and gains of ~20% compared to SR grants with a period of 1 TTI. The FTP throughput is similar to that of SPS grants.

The resource utilization is given in Figures 4 for DL and in Figure 9.1.4-5 for UL.

Table 9.1.4-2. Simulation parameters

Network	7 sites with 3-sector cells, 500m ISD
Bandwidth	10 MHz in all cells
Duplex	FDD
Traffic model	Single FTP file download
File size	1 MB
TCP initial window size	10
SR period	[1 TTI, 10 TTI]
UE arrival rate	Poisson Distribution, [2.82, 6.58, 9.3975, 12.22, 15.04] per second
Radio clock power consumption	10mW
RF RX mixer and LNA power consumption	72mW
RF RX filter and ADC power consumption per MHz	14mW/MHz
RX baseband power consumption (mW)	$25 + 7 \times \text{Mbps}_{\text{RX}}$
TX mixer power consumption	80mW
TX DAC power consumption per Mhz	16mW/MHz
TX baseband power consumption (mW)	$11 + 1 \times \text{Mbps}_{\text{TX}}$
Power consumption of PA (mW)	$72 + 17.5 P_{\text{TX}}^{0.784}$

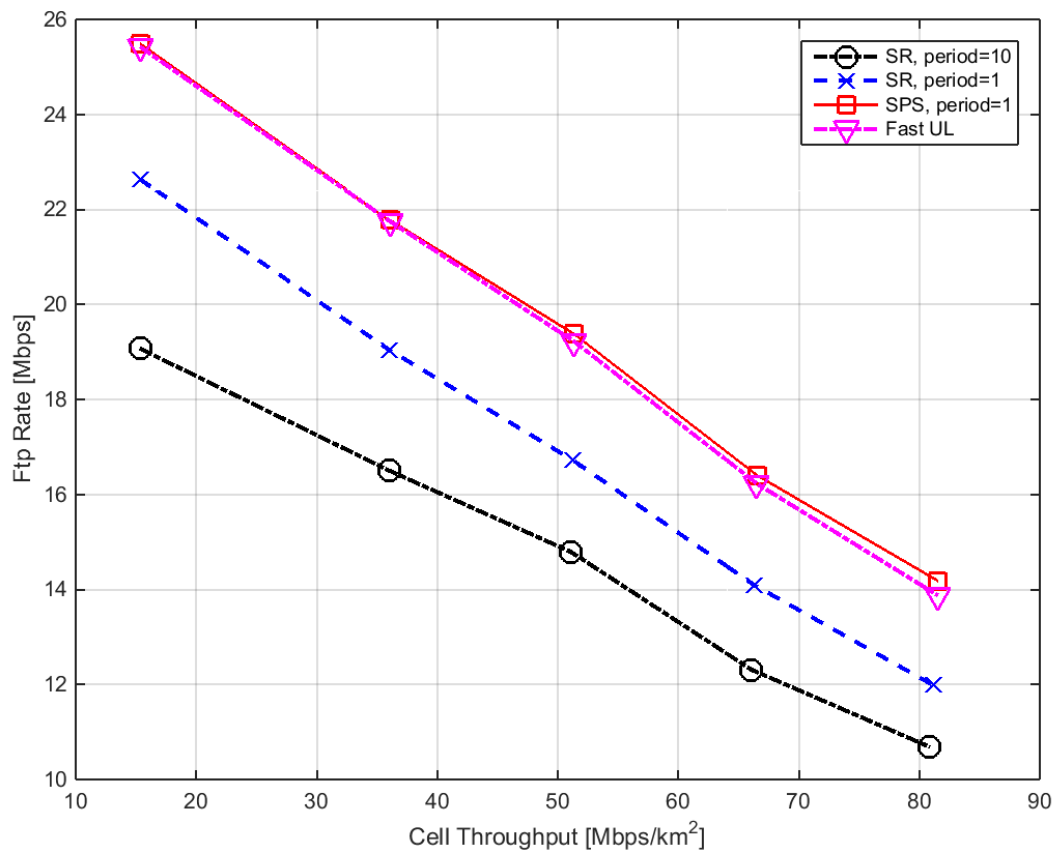


Figure 9.1.4-3. FTP Download Rate for fast UL grant (magenta) compared to SR (blue and black) and SPS (red)

In Figure 9.1.4-6 the energy consumption of the different grant mechanisms is shown. Here, we can see that the Fast UL grant solution exhibits the lowest energy consumption among the studied cases, about 10% lower than for SR grants. Compared to SPS, no padding is sent unless there is UL data, which saves battery. It should be noted that no DRX functionality has been assumed in these simulation. If DRX could be applied in the UE during the “inactive” phase with fast UL grant, additional reduction in power consumption can be expected.

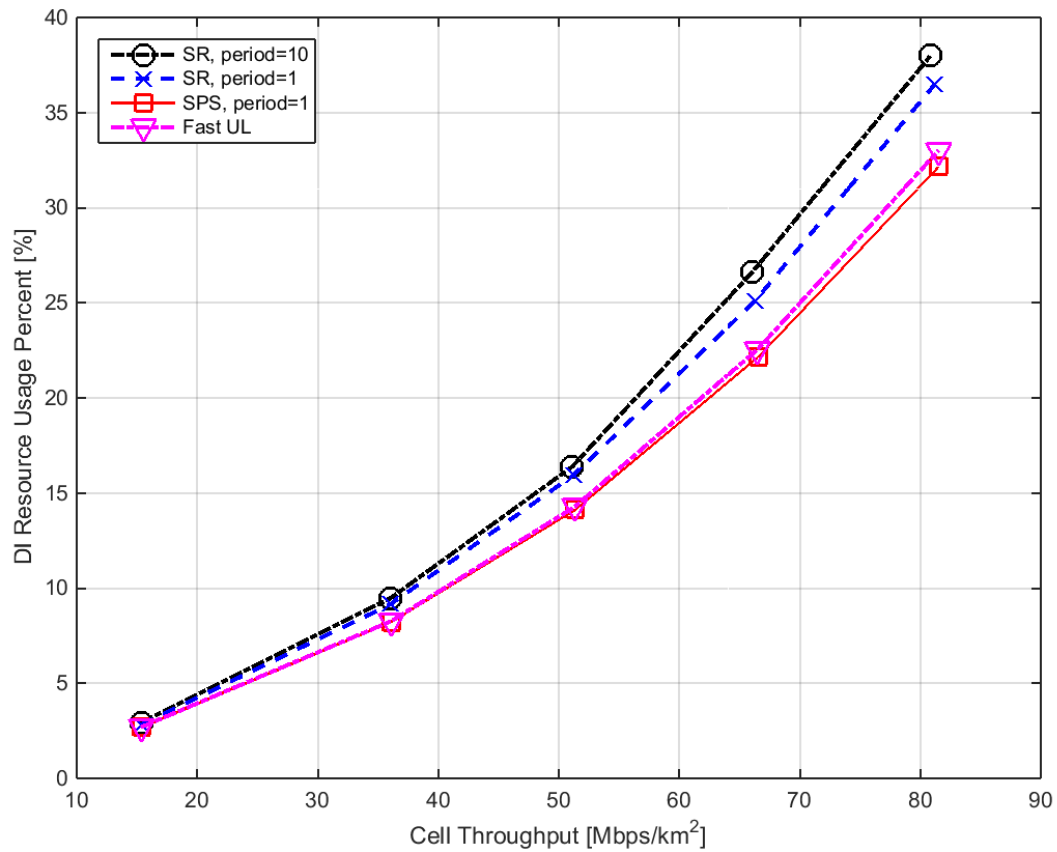


Figure 9.1.4-4: DL Resource Utilization for fast UL grant (magenta) compared to SR (blue and black) and SPS (red).

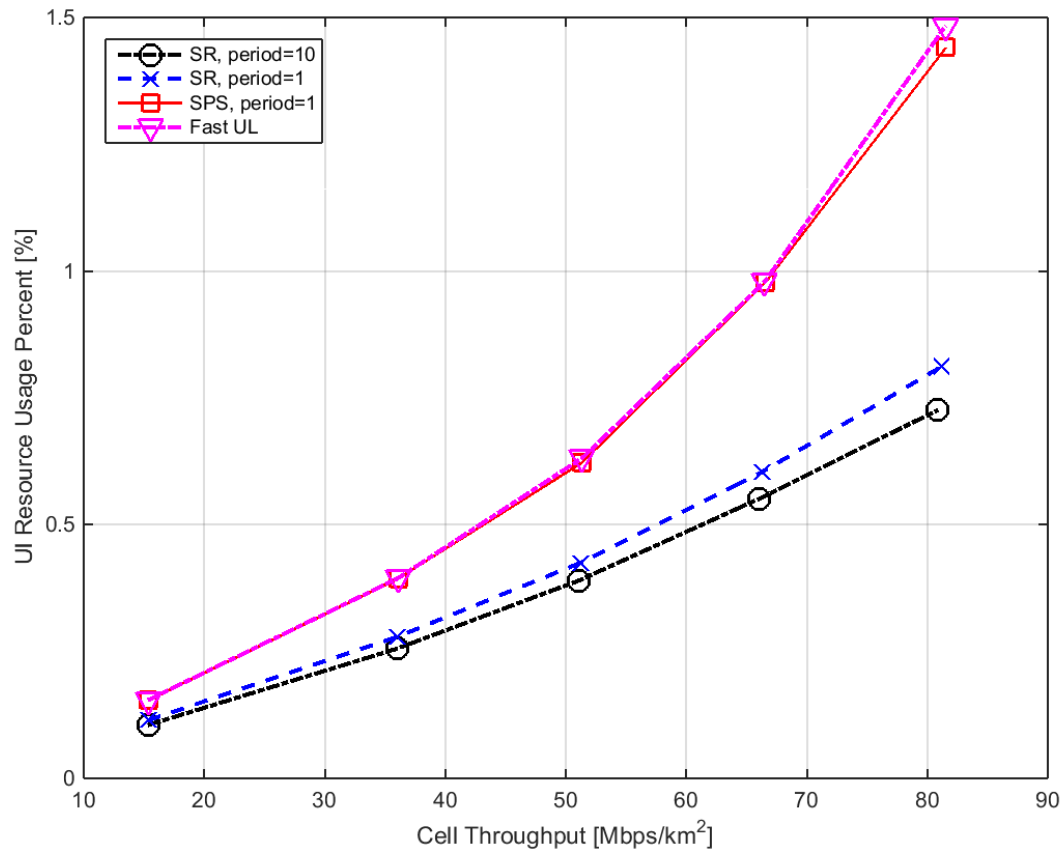


Figure 9.1.4-5: UL Resource Utilization for fast UL grant (magenta) compared to SR (blue and black) and SPS (red).

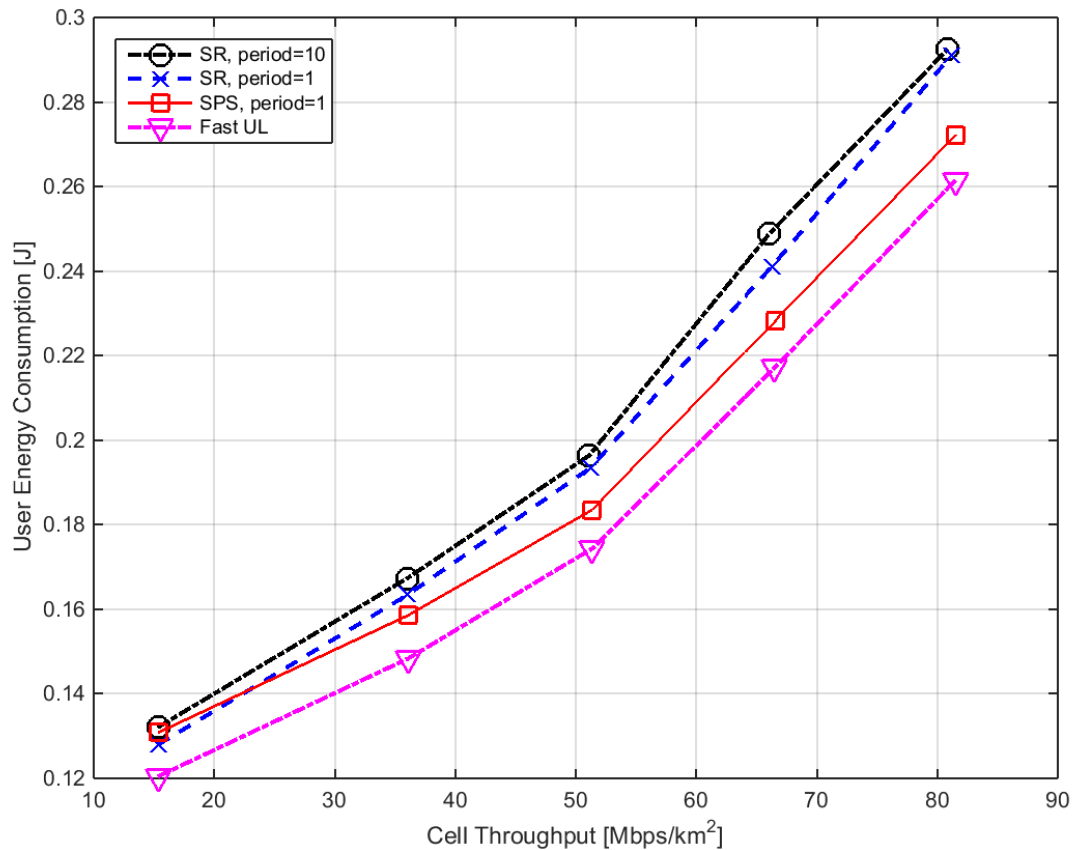


Figure 9.1.4-6. Energy consumption for fast UL grant (magenta) compared to SR (blue and black) and SPS (red).

Observation 12. Fast UL reduces the energy consumption compared to SR and SPS.

9.1.5 Simulation 5

System Performance with TTI shortening [6]

9.1.5.1 General information

For the service transmission, the whole latency includes backhaul latency (the transmission delay between eNB and application server) and RAN latency. Although the SI is only intended to shorten RAN latency, when we analysis the performance of RAN latency reduction, the impact of backhaul latency should not be neglected, and backhaul delay should be considered as an evaluation factor.

For FTP download application, since the improvement on the file download time brought by latency reduction would mainly on the TCP slow start phase, the improvement would be different and depend on the proportion of TCP slow start phase in the whole file download time. Hence, the downloaded file size should be considered as an evaluation factor.

Besides the file size, the throughput in Uu interface provided to UE (i.e. MCS, PRBs) also brings the impact on the benefit brought in the TCP slow start phase. Hence, the Uu throughput should also be considered as an evaluation factor.

As discussed in meeting, we should also take the L1 control and RS overhead into account.

For RAN latency, it includes data transmission time (i.e. TTI), RTT, and UL access latency (in UL direction, means the period from UL data arrival in UE to data transmission based on the receiving UL grant). Currently, TTI is 1ms; RTT is 8TTI (i.e. 8ms for FDD). UL access latency is about 14.5ms (in D-SR case). To reduce the latency further, for TTI, the smaller granularity should be evaluated, 1 OFDM symbol, 3 OFDM symbol and 7 OFDM symbols.

For extra UL latency, two existing solutions, i.e. Short D-SR and pre-allocation, with shorter UL latency are evaluated.

9.1.5.2 Simulation assumptions and parameters

The evaluation assumptions are given as below:

- 1 FTP download application is used in evaluation, and Downloading Response Time (DRT, sec) is regarded as the metric of performance evaluation. Wherein, DRT includes the time period from UE requesting FTP service to finishing downloading successfully;
- 2 Evaluation is based on single UE system with specific block error rate at air interface (i.e. 10% initial BLER);
- 3 Simulation parameters in Uu interface are given in table below.
 - For backward compatibility, we assume at least half of bandwidth should be reserved for legacy UEs;
 - To evaluate the impact on various backhaul latency, FTP file size and UL access, we consider the sufficient resource allocation with 50 PRBs and assume the good channel quality with MCS 20;
 - To evaluate the impact on the various Uu throughput, we also consider the limited resource allocation with 10 PRBs and assume middle channel quality with MCS 12.
- 4 L1 overhead (i.e. L1 control channel and RS) is considered as 10%, 20%, 30% and 50%, which is reflected into the TBS calculation. To unify the TBS calculation, we also use the same TBS calculation for the base line case (i.e. 1ms TTI), and the 20% L1 overhead is assumed. The detail on TBS calculation is given in Annex A1.5.
- 5 L2 overhead is not taken into account.

Table 9.1.5-1: Simulation parameters

Parameters	Assumptions
APP Server <--> eNB delay (ms)	0, 5, 10, 20, 30, 50
FTP File Size	10KByte, 100KByte, 1MByte, 2MByte, 5MByte
Uu throughput (Mbps)	27.907Mbps (MCS=20, PRB_Num=50) 2.848 (MCS=12, PRB Num=10)
TTI Duration	1ms (as baseline)
Available PRB number for LR UE	50 (for 20MHz BW)
Target Packet Error Rate	0.1
Max Transmission times	4
HARQ RTT	8*TTI

9.1.5.3 Performance results

9.1.5.3.1 Various backhaul latency

Assuming FTP file size 100KByte, Table 9.1.5-2 and Figure 9.1.5-1 give the improvement of each TTI length for various backhaul latency, with taking various L1 overhead (L1_OH) into account.

Table 9.1.5-2: backhaul latency

TTI length (in OFDM symbol number)	L1_OH		AppServer <-> eNB latency					
			0ms	5ms	10ms	20ms	30ms	50ms
14 symbol (baseline)	20% (Baseline)	DRT	0.183	0.263	0.342	0.500	0.660	0.977
7 symbol	10%	DRT	0.097	0.171	0.251	0.410	0.571	0.891
		Gain	46.99%	34.98%	26.61%	18.00%	13.48%	8.80%
	20%	DRT	0.102	0.176	0.255	0.413	0.574	0.894
		Gain	44.26%	33.08%	25.44%	17.40%	13.03%	8.50%
	30%	DRT	0.105	0.177	0.256	0.414	0.574	0.895
		Gain	42.62%	32.70%	25.15%	17.20%	13.03%	8.39%
	50%	DRT	0.120	0.186	0.258	0.416	0.577	0.896
		Gain	34.43%	29.28%	24.56%	16.80%	12.58%	8.29%
3 symbol	10%	DRT	0.058	0.124	0.203	0.363	0.523	0.842
		Gain	68.31%	52.85%	40.64%	27.40%	20.76%	13.82%
	20%	DRT	0.062	0.127	0.203	0.363	0.524	0.843
		Gain	66.12%	51.71%	40.64%	27.40%	20.61%	13.72%
	30%	DRT	0.067	0.130	0.203	0.364	0.524	0.843
		Gain	63.39%	50.57%	40.64%	27.20%	20.61%	13.72%
	50%	DRT	0.085	0.143	0.213	0.368	0.529	0.848
		Gain	53.55%	45.63%	37.72%	26.40%	19.85%	13.20%
1 symbol	10%	DRT	0.042	0.103	0.179	0.339	0.498	0.818
		Gain	77.05%	60.84%	47.66%	32.20%	24.55%	16.27%
	20%	DRT	0.047	0.106	0.179	0.340	0.499	0.819
		Gain	74.32%	59.70%	47.66%	32.00%	24.39%	16.17%
	30%	DRT	0.052	0.110	0.180	0.340	0.500	0.820
		Gain	71.58%	58.17%	47.37%	32.00%	24.24%	16.07%
	50%	DRT	0.071	0.124	0.192	0.344	0.503	0.823
		Gain	61.20%	52.85%	43.86%	31.20%	23.79%	15.76%
	Note: 1) DRT (sec) is the statistic metric, means downloading response time; 2) The gain is calculated by (1-DRT/Baseline_DRT), where, >0 means improvement, <0 means performance loss;							

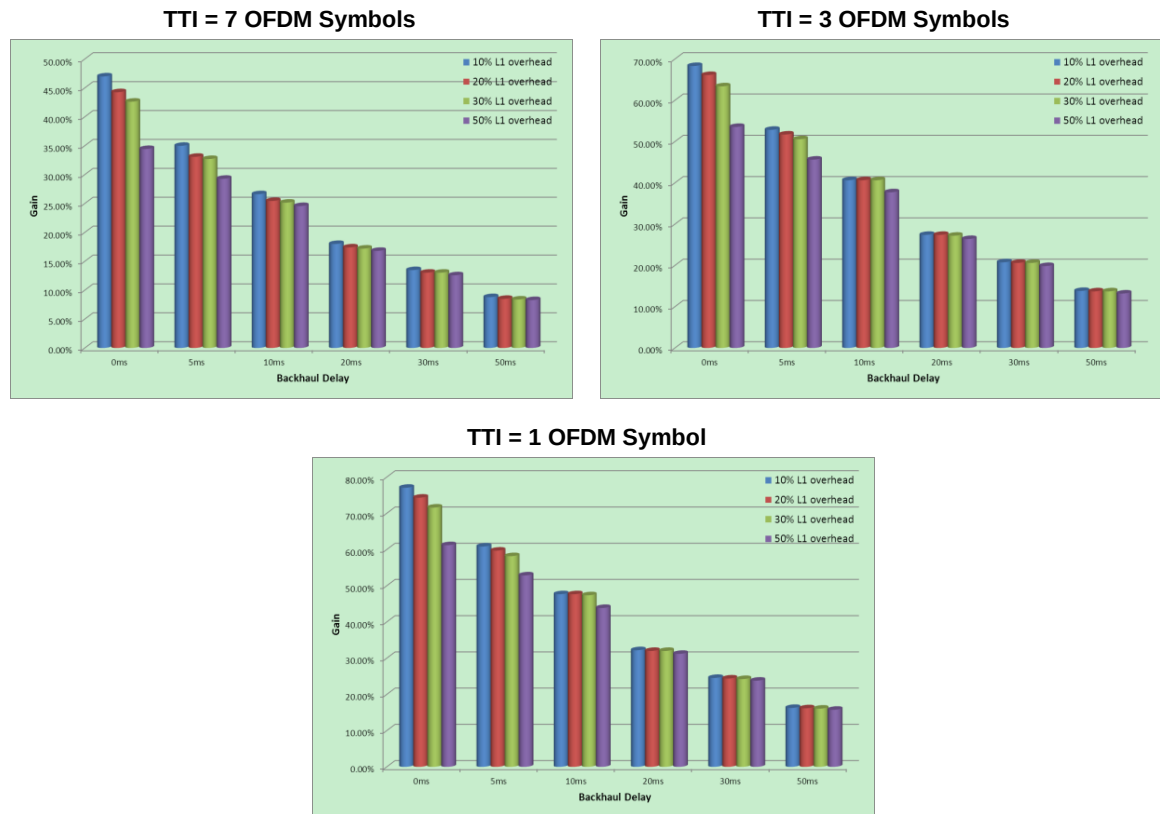


Figure 9.1.5-1: Gain for various backhaul latency

From Table 9.1.5-1, it can be seen that with the backhaul delay increasing the DRT difference of different TTI length is decreasing, and in case of backhaul delay 50ms, we can regard there is no much difference of different TTI length. And we think the reason is that TTI shortening only brings the improvement on U_u transmission, and the benefit is negligible compare to the long backhaul delay.

Observation 13. The longer the backhaul latency, the less the gain on DRT brought by shorter TTI.

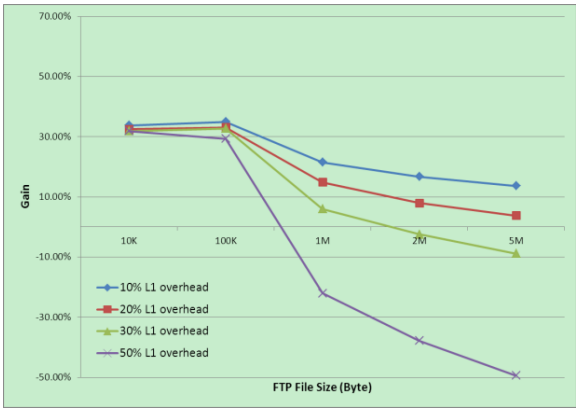
9.1.5.3.2 Various FTP file size

Assuming backhaul delay 5ms, Table 9.1.5-3 and Figure 9.5.1-2 gives the improvement of each TTI length for various FTP file size.

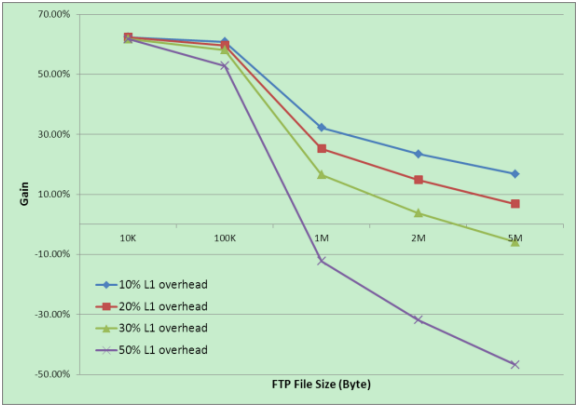
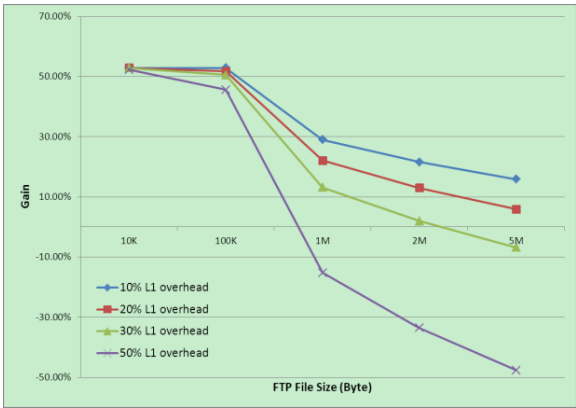
Table 9.1.5-3: Various file size

TTI length (in OFDM symbol number)	L1_OH		FTP File Size (Byte)				
			10K	100K	1M	2M	5M
14 symbol (baseline)	20%	DRT	0.157	0.263	0.634	1.036	2.263
7 symbol	10%	DRT	0.104	0.171	0.498	0.863	1.954
		Gain	33.76%	34.98%	21.45%	16.70%	13.65%
	20%	DRT	0.106	0.176	0.540	0.954	2.178
		Gain	32.48%	33.08%	14.83%	7.92%	3.76%
	30%	DRT	0.107	0.177	0.596	1.062	2.464
		Gain	31.85%	32.70%	5.99%	-2.51%	-8.88%
3 symbol	10%	DRT	0.107	0.186	0.774	1.428	3.381
		Gain	31.85%	29.28%	-22.08%	-37.84%	-49.40%
	20%	DRT	0.074	0.124	0.450	0.813	1.905
		Gain	52.87%	52.85%	29.02%	21.53%	15.82%
	30%	DRT	0.074	0.127	0.494	0.902	2.129
		Gain	52.87%	51.71%	22.08%	12.93%	5.92%
1 symbol	10%	DRT	0.074	0.130	0.551	1.016	2.417
		Gain	52.87%	50.57%	13.09%	1.93%	-6.81%
	20%	DRT	0.075	0.143	0.731	1.384	3.342
		Gain	52.23%	45.63%	-15.30%	-33.59%	-47.68%
	30%	DRT	0.059	0.103	0.430	0.793	1.883
		Gain	62.42%	60.84%	32.18%	23.46%	16.79%
1 symbol	10%	DRT	0.059	0.106	0.474	0.882	2.108
		Gain	62.42%	59.70%	25.24%	14.86%	6.85%
	20%	DRT	0.060	0.110	0.529	0.997	2.396
		Gain	61.78%	58.17%	16.56%	3.76%	-5.88%
	30%	DRT	0.060	0.124	0.712	1.366	3.323
		Gain	61.78%	52.85%	-12.30%	-31.85%	-46.84%

TTI = 7 OFDM Symbols



TTI = 3 OFDM Symbols



TTI = 1 OFDM Symbol

Figure 9.1.5-2: Gain for various file size

From Table 9.1.5-3, it can be seen that the gain becomes smaller even negative with the larger file size and higher L1 overhead: for 10Kbyte and 100Kbyte file size, there is significant gain with shorter TTI, and better in 10KB than in 100KB; however, with larger file size, the gain is decreased and even to negative due to higher L1 overhead. We think main factors, which lead to the performance decrease, are listed as below:

- 1 The throughput in Uu interface is significantly affected by physical control channel and RS overhead, especially for larger files, and lack of Uu capacity leads to longer latency in RAN side;
- 2 The latency reduction mainly speeds the TCP slow start phase. With file size increasing, the total download time is increasing, and the ratio of TCP slow start phase in entire download period would be smaller.

Observation 14. For FTP download application, the larger the download file size, the less (even negative) the gain on DRT brought by shorter TTI.

Observation 15. With the same TTI length, the higher the L1 overhead, the less (even negative) the gain on DRT.

9.1.5.3.3 Various Uu throughput

Assuming backhaul delay 5ms and FTP file size 100KByte, Table 9.1.5-4 gives the improvement of each TTI length for different Uu throughput.

Table 9.1.5-4: Various air interface condition

TTI length (in OFDM symbol number)	L1_OH		Uu Rate = 27.907Mbps MCS=20 PRB Num=50	Uu Rate = 2.848Mbps MCS=12 PRB Num=10
14 symbol (baseline)	20%	DRT	0.263	0.536
7 symbol	10%	DRT	0.171	0.438
		Gain	34.98%	18.28%
	20%	DRT	0.176	0.479
		Gain	33.08%	10.63%
	30%	DRT	0.177	0.536
		Gain	32.70%	0.00%
	50%	DRT	0.186	0.713
		Gain	29.28%	-33.02%
3 symbol	10%	DRT	0.124	0.409
		Gain	52.85%	23.69%
	20%	DRT	0.127	0.453
		Gain	51.71%	15.49%
	30%	DRT	0.13	0.509
		Gain	50.57%	5.04%
	50%	DRT	0.143	0.69
		Gain	45.63%	-28.73%
1 symbol	10%	DRT	0.103	0.397
		Gain	60.84%	25.93%
	20%	DRT	0.106	0.44
		Gain	59.70%	17.91%
	30%	DRT	0.11	0.497
		Gain	58.17%	7.28%
	50%	DRT	0.124	0.676
		Gain	52.85%	-26.12%

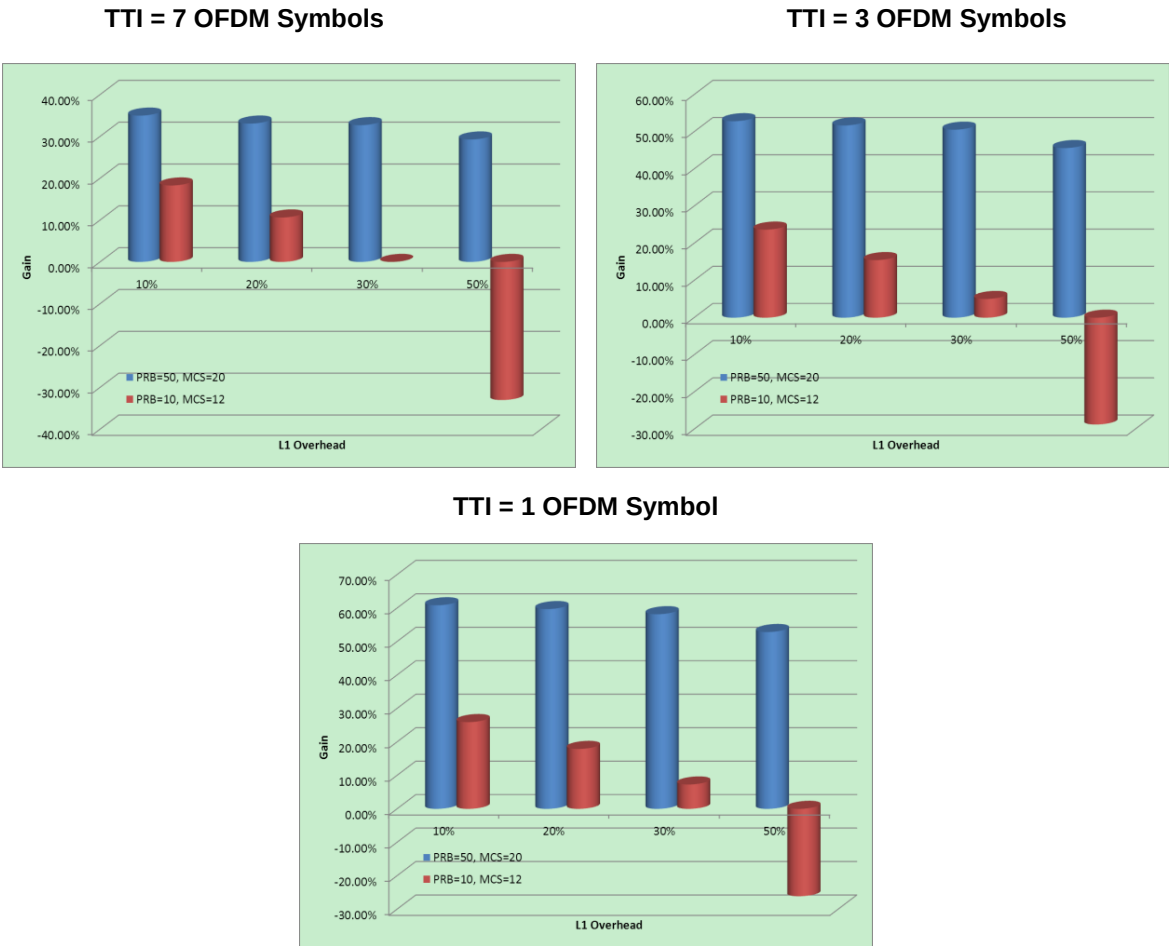


Figure 9.1.5-3: Various Uu Throughput

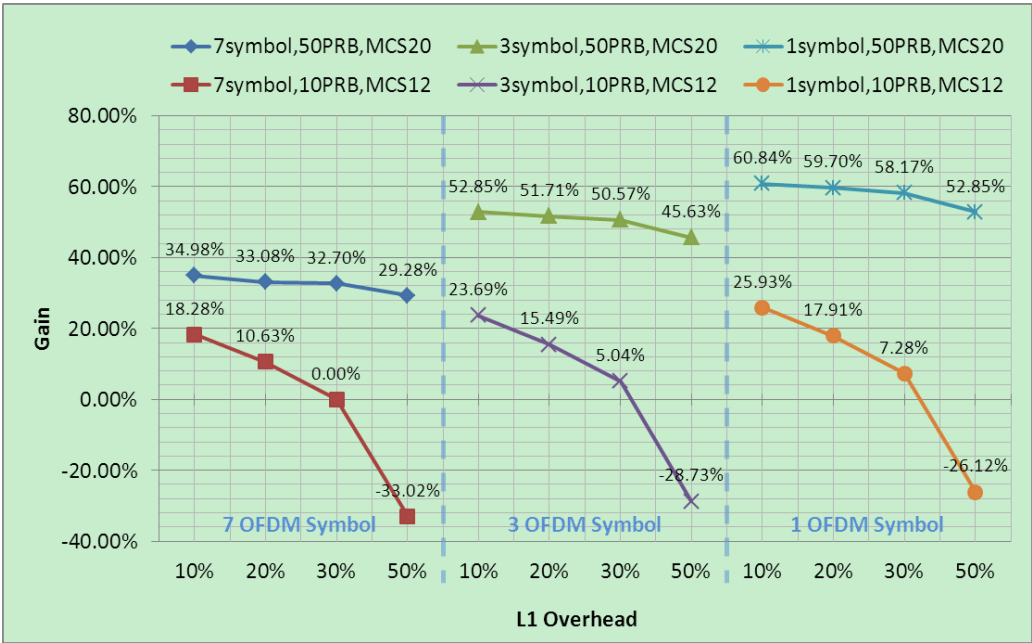


Figure 9.1.5-4: Various Uu Throughput

From Table 9.1.5-4 and Figure 9.1.5-3, it can be seen that the gain on DRT brought by shorter TTI is less when the provided Uu throughput is smaller; and in case of L1 overhead 50% and 1 symbol TTI, there is negative gain in case of

Uu throughput 2.848Mbps. We think that the main reason is that the smaller Uu throughput brings the bottleneck for FTP downloading.

Observation 16. The smaller the Uu throughput, the less the gain on DRT brought by shorter TTI.

Considering the network backward compatibility, for a cell, since the network resource is shared for legacy TTI and for shorten TTI, reserving more PRBs for shorten TTI would lead to the less resource for legacy TTI, which would bring bad impact on the throughput on legacy UEs. In addition, since there are many UEs in one cell, allocating resource with 50 PRBs to a UE in every TTI can bring the low latency to that UE, but the UE number supported in that cell would be dramatically decreased. Hence, it is not reasonable to allocate resource with 50 PRBs to each UE each TTI. Comparing to that, the resource allocation with 10 PRB is more reasonable.

From Figure 9.1.5-4, it can be seen that in case of smaller Uu throughput, considering the same L1 overhead, there are not so much difference on the gain for each shorter TTI (i.e. 10.63% Vs. 15.49% Vs. 17.91%). In addition, Since the existence of RS signal and L1 control channel are inevitable for efficient data transmission, the shorter TTI would bring more L1 overhead, when we compare the gain of each TTI length, it is not appropriate to compare the gain of the same L1 overhead in different TTI length. For example, when evaluating the gain brought by 7 symbol and 3 symbol TTI length, it is more reasonable to compare the gain in (7 symbols, 20% L1 overhead) and in (3 symbols, at least 30% L1 overhead), 10.63% Vs. 5.04%, and 7 symbol TTI length is a little better than 3 symbol TTI length.

Observation 17. When evaluating and comparing the gain of different TTI length, it is reasonable to assume shorter TTI length leads to more L1 overhead.

Observation 18. In case of small Uu throughput, the gain brought by 20% L1 overhead and 7 symbols TTI is more than by 30% L1 overhead and 3 symbols TTI.

9.1.5.3.4 Fast UL access

UL access latency also brings impact on downloading performance, i.e. TCP ACK will feedback slowly if UL access latency is too big, which would impact the TCP transmission window increasing.

In Rel-10, 3GPP already discussed the latency reduction topic and at least the short D-SR and Pre-allocation, as we called fast UL access solutions, were supported by standardization or implementation, here we evaluated the performance gain with different UL access means (i.e. normal D-SR, short D-SR, and pre-allocation), and also with different TTI length.

Assuming download file size 100KByte and backhaul delay 5ms, Table 9.1.5-5 gives the improvement of various UL access solutions in different TTI length. In Table 9.1.5-5, gain (C) is the gain of the various UL access solution combined with TTI shortening, compared with the normal D-SR in 14-symbol TTI length (**the value in red**); and gain (UL) is the individual gain for various UL access solution, compared with the normal D-SR in each TTI length.

Note: the UL access latency components for UL access solutions are stated as in Annex A1.5.

Table 9.1.5-5: Various UL latency combining with TTI shortening

UL Access		TTI Length (in OFDM symbol number)			
		14 symbol	7 symbol	3 symbol	1 symbol
Normal D-SR (D-SR Interval = 5TTI)	DRT	0.263	0.176	0.127	0.106
	gain (C)	0.00%	33.08%	51.71%	59.70%
	gain(UL)	0.00%	0.00%	0.00%	0.00%
Short D-SR (D-SR Interval = 1TTI)	DRT	0.244	0.168	0.124	0.105
	gain (C)	7.22%	36.12%	52.85%	60.08%
	gain(UL)	7.22%	4.55%	2.36%	0.94%
Pre-allocation (UL grant Interval = 1TTI)	DRT	0.219	0.155	0.119	0.103
	gain (C)	16.73%	41.06%	54.75%	60.84%
	gain(UL)	16.73%	11.93%	6.30%	2.83%

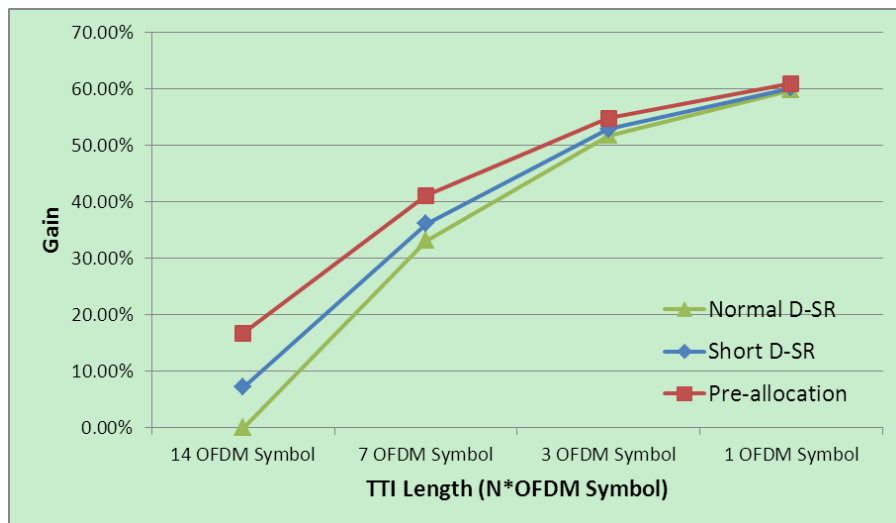


Figure 9.1.5-5: Gain(C) for various UL access solutions

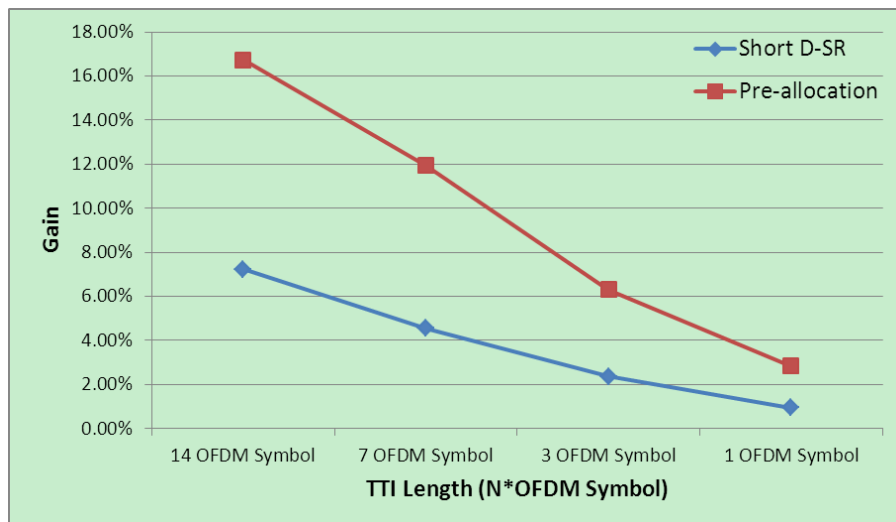


Figure 9.1.5-6: Gain (UL) for various UL access solutions

Figure 9.1.5-5 gives the gain for the combination of TTI shortening and UL latency reduction, and based on that it can be observed that compared with legacy UL access solution (i.e. 1ms TTI and 5ms D-SR period) the DRT would be dramatically decreased when the TTI and UL latency is shortening. Compared with the individual gain for TTI shortening given in Figure 9.1.5-4 (in case of PRB_Num=50, MCS=20, L1 OH=20%), TTI shortening gives most contribution.

Observation 19. Shorter TTI and shorter UL latency leads to shorter DRT, and most contribution is given from TTI shortening.

Figure 9.1.5-6 gives the individual gain for UL latency reduction solutions in each TTI length, respectively. It can be seen that the shorter the UL access latency is, the more the gain is; the shorter the TTI is, the more the gain is as well. However, we can't see an obvious gain using fast UL access solutions in shorter TTI length, e.g. 3-symbol and 1-symbol TTI, as we see in normal TTI length. We think the reason is that faster TCP ACK feedback in uplink direction will speed up the TCP slow start, which will make the downloading time decreasing. On the other hand, shorter TTI has already speed up the UL access; so with shorter TTI, we can't see an obvious gain using fast UL access solutions as we see in normal TTI length.

Observation 20. With the same TTI length, the shorter the UL access latency, the shorter the DRT.

Observation 21. With TTI shortening, the performance gain brought by fast UL access decreased.

9.1.6 Simulation 6

Evaluation results for TTI reduction [6]

9.1.6.1 Simulation assumptions

When evaluating the gains provided by one symbol length TTI, we will consider various losses in L1 data rate due to the additional overhead caused by reduced TTI length. In Annex A1.3, we provide the estimated L1 data rate loss for one symbol length TTI. Table A1.3-1 assumes that the DL channel estimation is based on CRS, and Table A1.3-2 assumes that the DL channel estimation is based on DM-RS. From Table A1.3-1 and Table A1.3-2, we can see that the loss in L1 data rate is vary from -1.21% to 30.48%, hence we will use the L1 data rate loss of 0%, 5%, 10%, 20% and 30% respectively in the system simulation.

The simulation assumptions on radio aspects, TCP parameters and FTP traffic model are provided in Annex A1.2.

This simulation will not consider the increased transmission efficiency in high-speed scenarios due to the fast channel sensing (i.e. fast CSI report caused by shortened TTI).

Simulation results on UPT gain

Figure 9.1.6-1 illustrates the average UPT gain provided by one symbol length TTI for different L1 data rate loss, when the FTP file size is 0.5Mbytes. The number of UEs per cell assumed in the simulation is 1, 5, 10, 20 and 35, which roughly corresponds to the system load of 3%, 12%, 25%, 50% and 87% respectively.

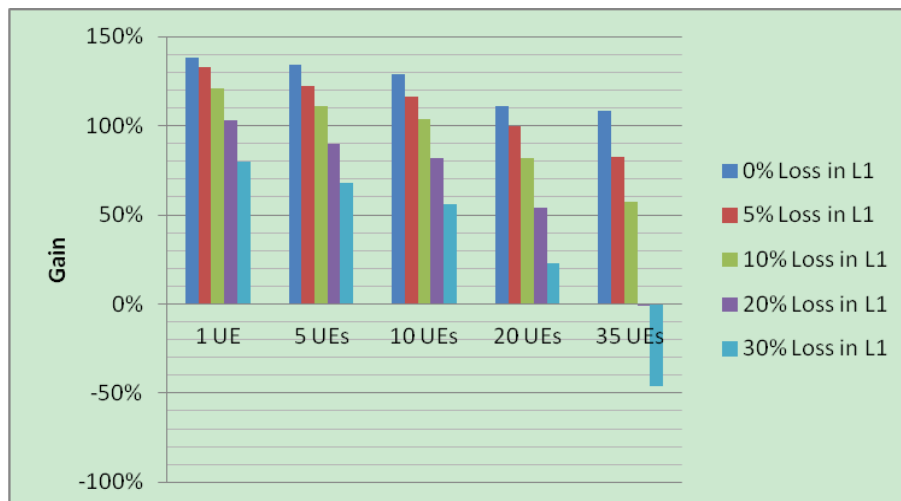


Figure 9.1.6-1: Average UPT gain provided by one symbol length TTI, for different L1 data rate loss

From Figure 9.1.6-1, it could be observed that one symbol length TTI can bring significant UPT gain in general, and more specially:

- The UPT gain gradually decreases when the loss in L1 data rate increases, as the available PDSCH capacity for data transmission is decreased. Nevertheless, even when the loss in L1 data rate reaches 30%, the UPT gain is still considerable, e.g. beyond 20%.
- The UPT gain gradually decreases when the system load increases (i.e. when there are more UEs in the cell), this is because the UPT gain provided by shortened TCP layer RTT will be compromised by the decreased transmission/reception data rate of the UE. In case the system is heavily loaded (i.e. 35 UEs per cell), one symbol length TTI only brings marginal UPT gain even negative UPT gain.

Figure 9.1.6-2 illustrates the 5% UPT gain provided by one symbol length TTI for different L1 data rate loss, when the FTP file size is 0.5Mbytes.

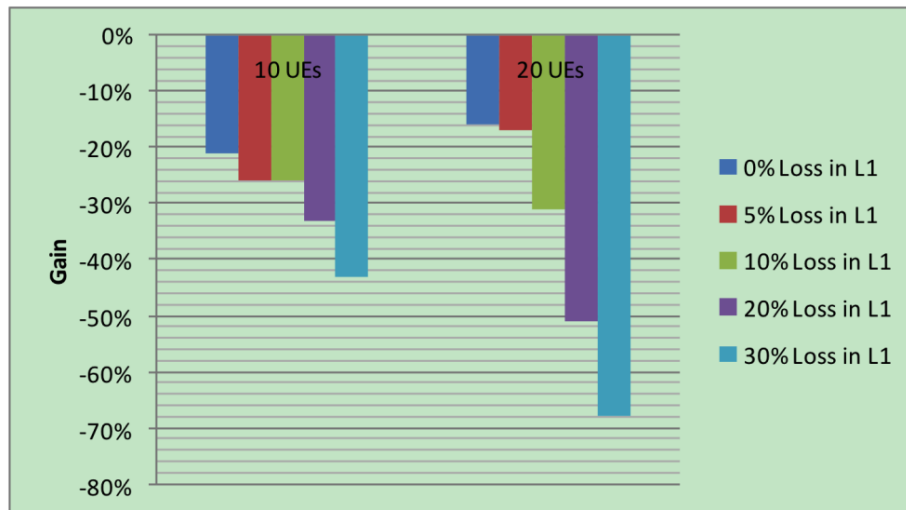


Figure 9.1.6-2: 5% UPT gain provided by one symbol length TTI, for different L1 data rate loss

From Figure 9.1.6-2, it could be observed that the 5% UPT gain provided by one symbol length TTI is negative, even if we don't consider any L1 data rate loss (i.e. 0% loss) in the simulation. This is because for 5% of the UEs, only a few bytes could be accommodated in each TTI given that the TTI is only one symbol length and the UE is in poor channel condition. Small TB size in each TTI will cause more segments for packet delivery, and consequently will bring significant MAC and RLC overhead. In addition, the 5% of UEs might suffer from power limitation, which may prevent timely report of the TCP ACK, and this will enlarge the TCP layer RTT hence further decrease the UE throughput. For the L1 data rate loss of 30%, the negative UPT gain for 5% of the UEs is significant, i.e. beyond 40%.

Observation 22. One symbol length TTI can bring significant UPT gain in average, even if the loss in L1 data rate reaches 30%.

Observation 23. If the system is high loaded, or if the UE is in cell edge, one symbol length TTI may bring negative UPT gain.

Figure 9.1.6-3 illustrates the average UPT gain provided by one symbol length TTI for different FTP file sizes, assuming the L1 data rate loss is 0% and the number of UEs per cell is one.

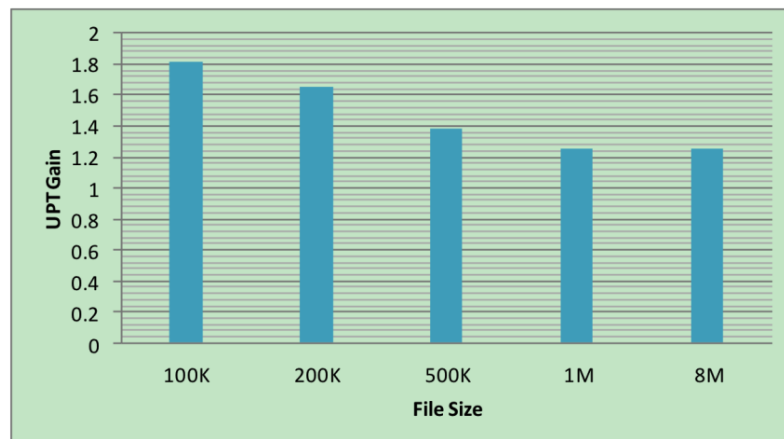


Figure 9.1.6-3: UPT gain provided by one symbol length TTI, for different file sizes

From Figure 9.1.6-3, it could be observed that the average UPT gain provided by one symbol length TTI decreases as the FTP file size increases. This is because the accelerated TCP slow start phase by shortened TTI is the major contributor for the UPT gain, and for small FTP file the TCP show start phase takes a big percentage of the total FTP download time while this not the case for large FTP file. Nevertheless, the average UPT gain is still considerable when the FTP file size arrives at 8Mbytes, which is about 25%. This is because the UPT will still benefit from the shortened TCP layer RTT caused by one symbol length TTI, unless the system is high loaded.

Observation 24. One symbol length TTI can also bring significant UPT gain in average for large file size.

9.1.7 Simulation 7

Performance evaluation of latency reduction enhancements [6]

9.1.7.1 Simulation setup

This section analyses the performance impact of shorter TTI and fast UL. System simulations have been conducted to compare different approaches for latency reduction. The simulations are run in the 3GPP macro scenario, i.e. 3GPP Case 1 macro scenario with 7 sites with 21 cells in total and wrap-around. For the UEs, we use a quasi-static model with static UE mobility and TU channel model with 3 km/h speed.

The performance evaluation is done for FTP traffic model 1 [10], where the load is varied by having different UE arrival rates for a fixed 0.5 MB file size: 26.25, 52.5, 78.75, 105 dropped per second to whole network, which translates to an offered load of 5-20 Mbps per cell. Additionally we also consider the impact of FTP file size. 0.1 MB and 2 MB file sizes are simulated as well (this affects the UE arrival rate proportionally). In the simulations the FTP traffic is transported using TCP. This enables observing the impact of TCP slow start phase, which is where the reduced latency is particularly beneficial. After the file has been transmitted, the UE is removed from the system (so new files are generated for new users). So it is assumed that each new file transfer starts a new TCP connection with slow start.

Additional L1 and L2 overhead due to the shorter TTI is modelled so that in the reference case (14 symbol TTI) the overhead is assumed to be 20% and we have additional cases with 30%, 40%, 50% overhead. We range several different values, as it is not yet clear what would be a realistic assumption here.

More details of simulation assumptions can be found in table 9.1.7-1. Section 9.7.1.2 describes the simulated schemes and Section 9.7.1.3 and 9.7.1.3 present the performance evaluation results.

Table 9.1.7-1 Simulation parameters

Parameter	Assumption
System bandwidth	20 MHz
Duplex mode	FDD
Carrier frequency	2GHz
Cell layout	Hexagonal grid, 7 sites, 21 cells per site, with wrap-around
Number of UEs	According to offered load
Inter-site distance	500m
UE speed	3 km/h, quasi-static model
Antenna configuration	15 MRC
eNB TX power	46 dBm
UE Tx power	23dBm
Channel model	Typical Urban
Pathloss model	25.814
Lognormal shadowing, std. dev.	8dB
Penetration loss	20dB
CQI measurement period	5 ms
SRS reporting period	5 ms
SR to grant	8ms – scaled down with shorter TTI
HARQ RTT	8ms – scaled down with shorter TTI
SR Period	Varied 1,5,10,20 ms
DRX	OFF
Initial TCP Window	3 x 1500 Bytes (MSS), RFC 5681, section 3.1
Initial Ssthresh	150 x 1500 Bytes (MSS)
Ssthresh	Dynamic according to RFC 5681 [11], sections 3.1 and 3.2
FTP file size	0.5 MB
User arrival rate λ	FTP model 1 with UE arrival according to Poisson process [10]
Scheduler	TD: PF, FD: PF
Maximum number of scheduled users per TTI	10 (max)
L1 overhead	Varied (ref (20%), 10, 20, 30 %-units increase in overhead)
Core network delay	2ms, 10ms
Load per macro cell	5, 10, 15, 20 Mbps
TTI Length	Varied (2symbols, 7 symbols, 14 symbols)

9.1.7.2 Simulated schemes

- Shorter TTI (7 symbols or 2 symbols)

Shorter TTI approach refers to shortening the TTI length from the currently specified 14 symbols and 1 ms to 7 symbols i.e. 0.5 ms, or even going down to a 2 symbol TTI i.e. 0.14 ms. Having shorter length TTI and high data rates would imply that more PDUs need to be handled with higher L1 and L2 overhead, which can at least to some degree offset the benefits of the reduced round trip time.

- Fast UL grant (i.e. pre-scheduling)

In the simulations the UE is assumed to always have resource granted (in practice this would be when the eNB expects some UL data from the UE, e.g. TCP ACK). Fast uplink scheduling can be achieved with protocol enhancements in L2 like in MAC protocol without changing physical layer principles like TTI length. However, it is applicable not only with the current 14 symbol TTI, but also with shorter TTI lengths.

- UL scheduling based on SR (as presented in Figure 1)

SR based UL grant refers to the (baseline) approach of UE sending SR in order to get an UL grant. This scheme was used as a reference for the other cases.

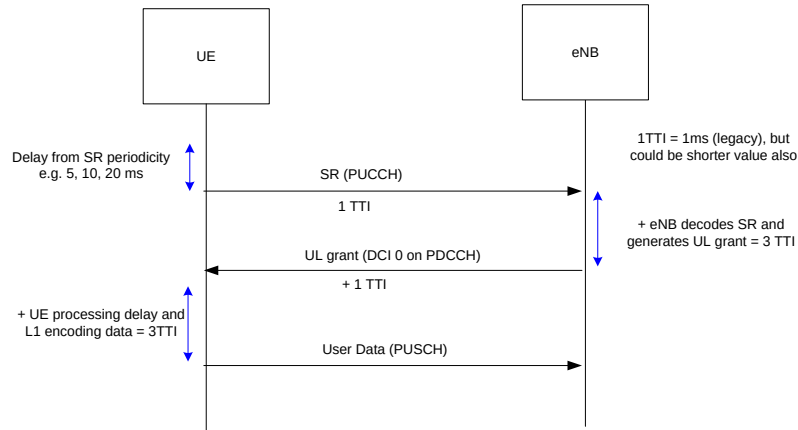


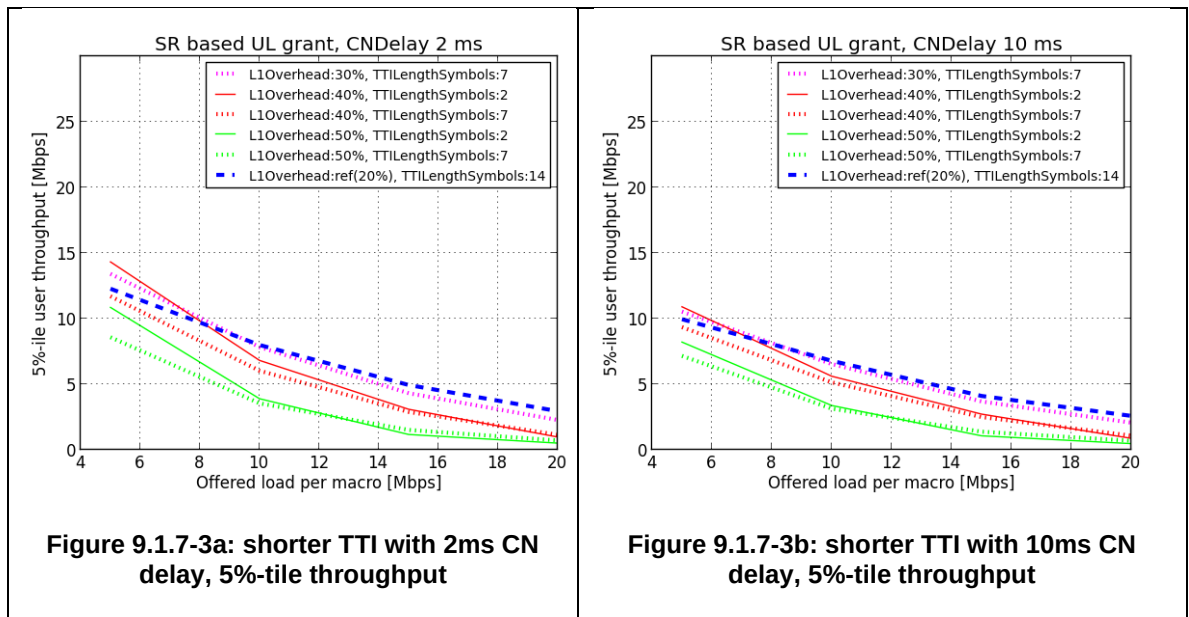
Figure 9.1.7-2 SR based UL grant procedure

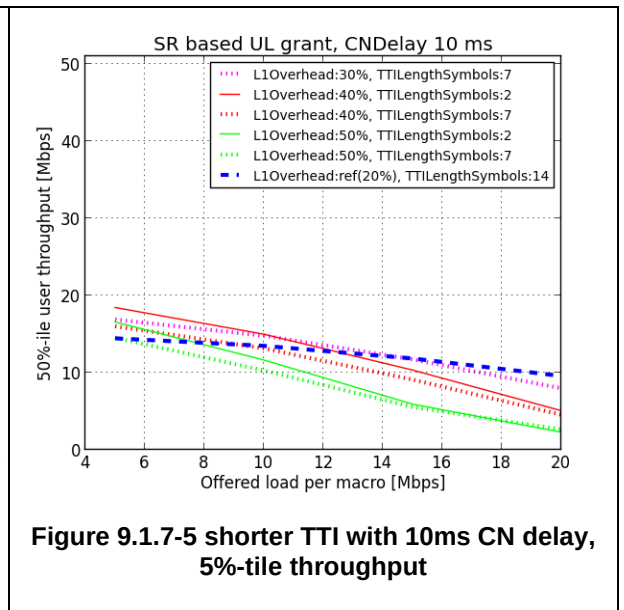
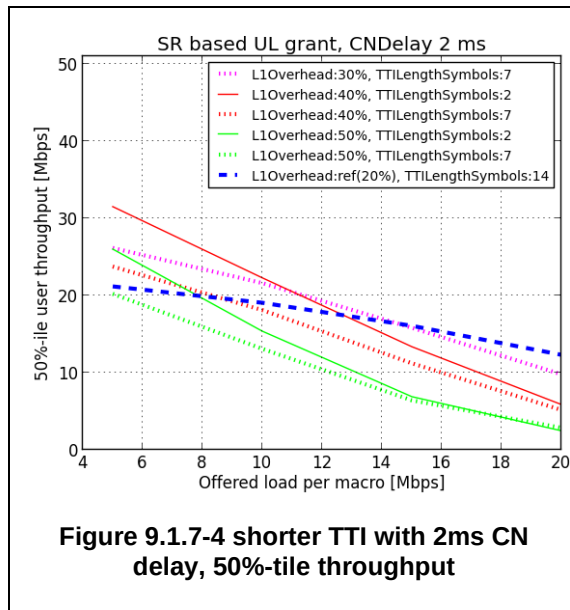
9.1.7.3 Simulation results for shorter TTI

The performance metric used for comparing the different schemes is the user throughput measured over the entire FTP file transfer. This is inversely proportional to the total delay of the file.

A: Effect of shorter TTI

Performance with 2 and 7 symbol TTI compared to the 14 symbol TTI for different L1 overhead, and core network delay (CN delay) assumptions. The results are presented for SR based UL grant (i.e. without pre-scheduling enhancement).





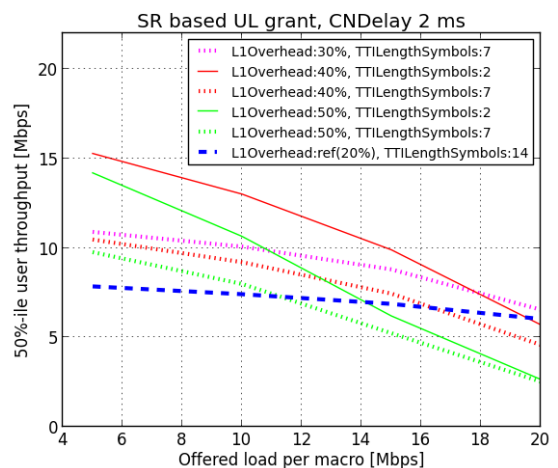
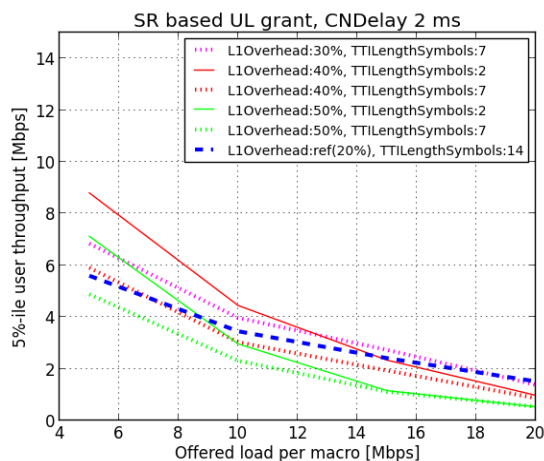
As indicated by the results, the potential gain from shorter TTI depends on how much L1/L2 overhead is assumed, the load of the cell as well as the CN delay. With L1/L2 overhead increased by 10 %-units, the difference between 2 symbol TTI and 7 symbol TTI is visible in low load situation. Increasing the load or the L1/L2 overhead the performance starts to deteriorate with shorter TTI compared to reference case. With higher load, a UE can no longer be scheduled in every (short) TTI, but may have to wait its turn. This reduces the latency benefit from the shorter TTI.

Observation 25. The potential gain from having shorter TTI depends on how much L1/L2 overhead is assumed and the load of the cell. It would be beneficial to control when shorter TTI are being used.

Observation 26. The difference between 2 symbol TTI and 7 symbol TTI is visible only in low load situation.

Observation 27. Increasing the load the performance starts to deteriorate resulting in negative gain with shorter TTI when the effect of scheduling delay becomes visible.

A: Effect of shorter TTI: comparing different file sizes



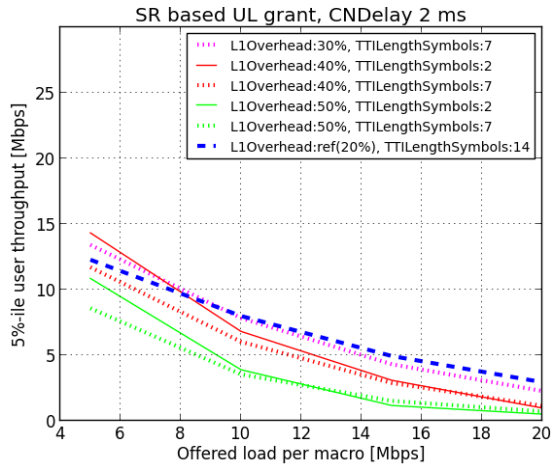


Figure 9.1.7-8 500kB file size, 5%-ile UE throughput

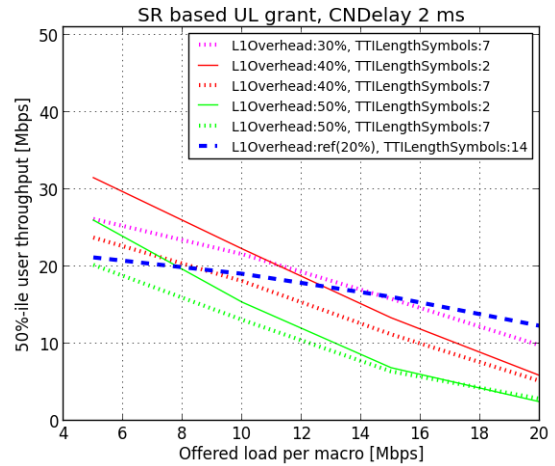


Figure 9.1.7-9 500kB file size, 50%-ile UE throughput

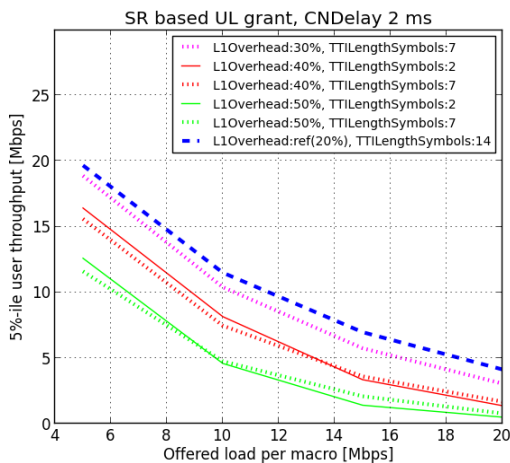


Figure 9.1.7-10 2MB file size, 5%-ile UE throughput

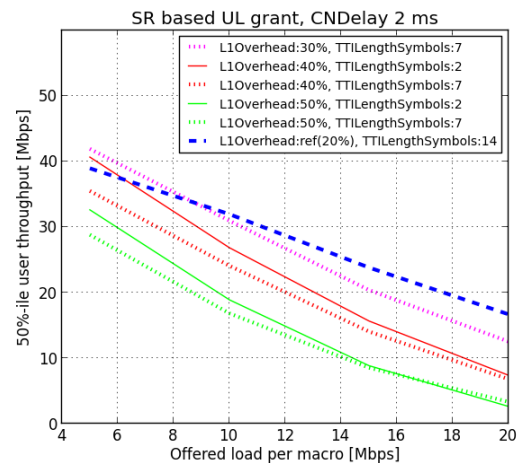


Figure 9.1.7-11 2MB file size, 50%-ile UE throughput

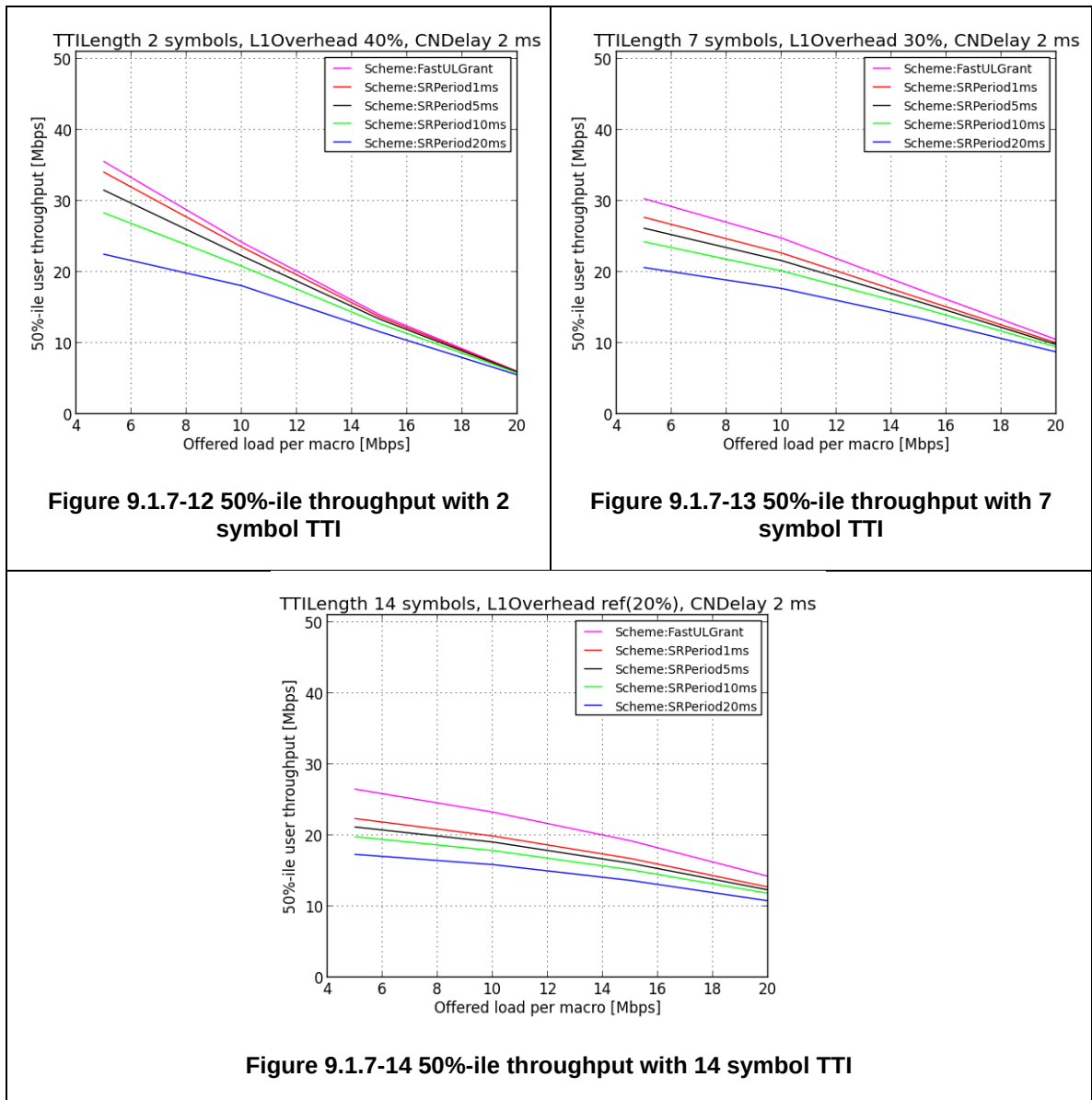
With smaller file size (100kB), the shorter TTI gives more gain due to the TCP slow start phase being more emphasized. With increased file size (2 MB) the benefits on the other hand start to diminish and additional overhead dominates. Similar observations are valid for both 5%-ile and 50%-ile UE throughput.

Observation 28. The potential gain of shorter TTI depends on file size. The benefits of shorter TTI diminish with larger file size. It would be beneficial to control when latency-reducing techniques are being used.

9.1.7.4 Simulation results for Fast UL grant with shorter TTI

‘SR based UL grant’ is used as reference scheme for the ‘Fast UL grant’.

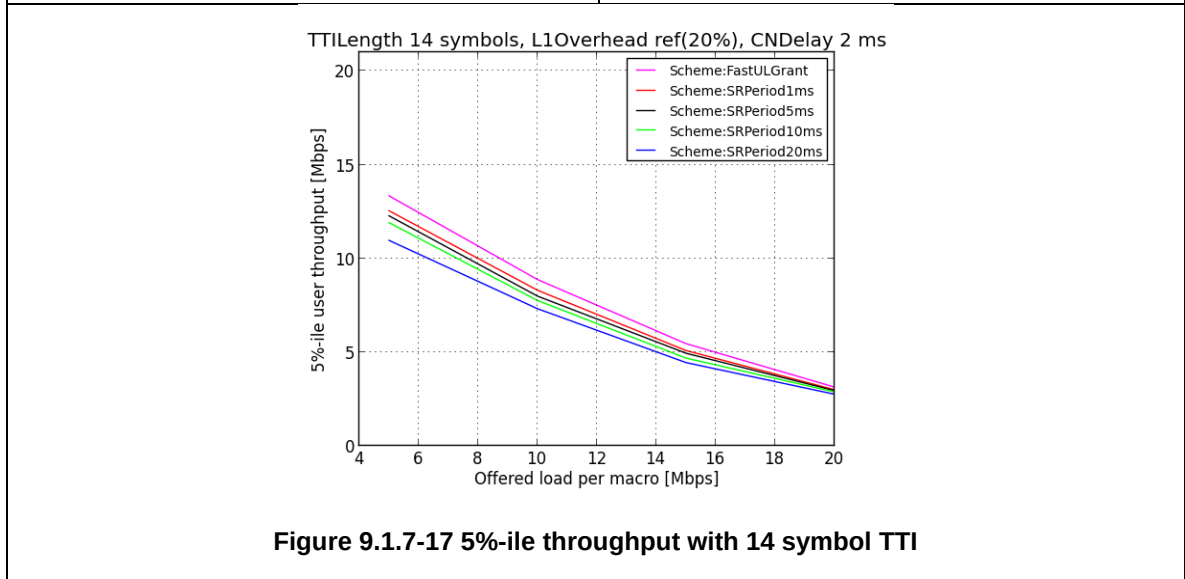
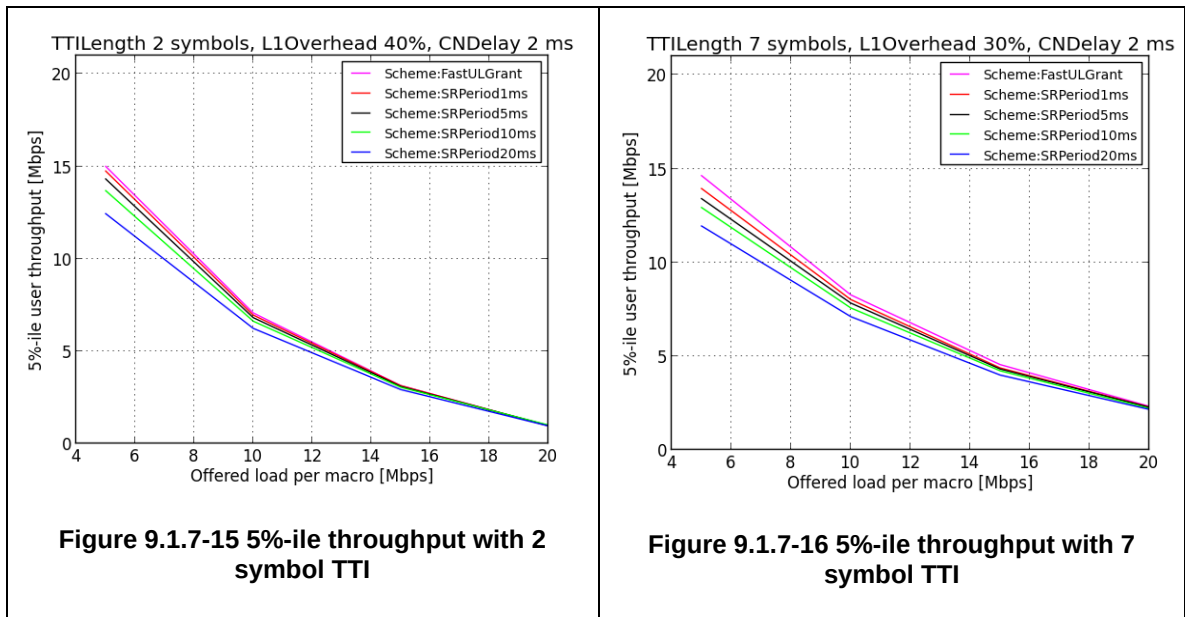
A: Comparison of ‘Fast UL grant’ vs. ‘SR based UL grant’



Observation 29. Fast UL grant approach can give a significant performance improvement over (legacy) SR based UL grant. This is especially the case when longer SR periodicities are used.

Observation 30. Pre-scheduling is beneficial over SR based UL grant with shorter TTI as well.

Similar observations can be seen for 5%-ile UE throughput.



9.1.8 Simulation 8

TTI reduction gain with additional L1/L2 overhead [6]

9.1.8.1 Simulation assumptions

In this evaluation, it is assumed that the core network and the Internet delay is very small and is independent of the TTI values. The assumed latency between eNB and TCP Server is 1-2 ms.

The average DL data-rate is kept the same for different TTIs. This is achieved by scaling the TBS sizes according to the duration of TTI compared to the current 1ms TTI.

For the FTP based file download using TCP protocol, TCP New Reno is used.

User perceived throughput (UPT) is calculated as total TCP layer data downloaded divided by the time elapsed. Initial time for setting up the network and for the UE to attach to network is removed as offset from the results such that the FTP request is sent to the server from the UE at time 0.

A saturation data-rate of approximately 21Mbps (example scenario is 3 CA 20MHz each, cell spectral efficiency ~3.5bps/Hz, 10 UEs/cell) is assumed. MCS index of UE is set such that on an average this data-rate is achieved when TCP has fully ramped up (channel adaptation is disabled).

In addition, for this study, fixed additional L1/L2 overheads due to reduced TTI handling are assumed. The additional overheads are simulated by scaling down the TBS accordingly during scheduling, both in DL and in UL. Because of the additional overhead, TCP “saturation” data-rate will be different for different overhead values, so ramp-up time taken to reach 90% saturation rate is not a fair metric for comparison across different overhead values within the same TTI. Therefore, FTP file download time (defined as the time duration between the instant when UE sends the FTP request and when the file completes download) and user perceived throughput (UPT) during the entire file download (total file size divided by file download time) is considered. Different file sizes for download are considered.

9.1.8.2 Evaluation results

Figure 9.1.8-1 and Table 9.1.8-1 show the file download time and user perceived throughput considering different additional L1/L2 overheads for different TTI. 1ms TTI with no additional L1/L2 overhead is taken as baseline for comparison. Various file sizes are considered to cover the cases of typically small, medium and large file downloads (100KB, 500KB and 1MB).

It is observed that for small FTP file sizes (100KB), even with 40% additional L1/L2 overhead, 0.5ms and 0.1ms TTI significantly outperform the baseline 1ms TTI – for 0.5ms the gain is about 43% increase in UPT (decrease in file download time) and for 0.1ms, the UPT is doubled. This is because the system is not overloaded during the initial ramp-up time (as sender TCP has not picked up yet), thus additional overhead does not affect the initial part of the TCP slow-start for the same TTI value. However, note that the slow-start behavior is different for different TTI (as shown in earlier results). Therefore, the gain due to improvement in slow-start behavior outperforms the loss due to additional L1/L2 overhead.

Observation 31. For smaller FTP download using TCP, the user perceived throughput improvement from the TCP slow-start outperforms the potential degradation even with additional L1/L2 overhead in the shorter TTI.

For medium file size (500KB), similar UPT compared to 1ms baseline TTI is observed with 20-30% additional overhead in 0.5ms TTI and even when more than 30% additional L1/L2 overhead is present in 0.1ms TTI.

When FTP file size is further increased, as expected, the 1ms TTI baseline scenario also can obtain high UPT. In such case, the amount of additional overhead due to reduced TTI without losing UPT performance is lower. For example, for 1MB file size, it is observed that similar UPT as in 1ms TTI is obtained with 10-20% additional overhead in 0.5ms and 0.1ms TTI.

Observation 32. For higher size FTP download using TCP, the user perceived throughput may be degraded in the shorter TTI if additional L1/L2 overhead is high.

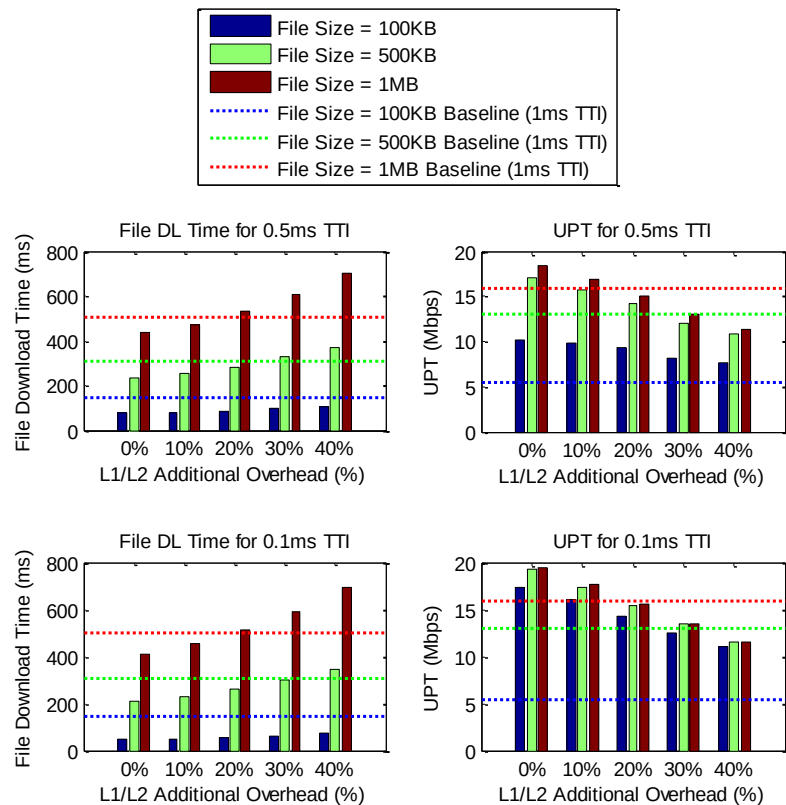


Figure 9.1.8-1 FTP file download performance with TTI reduction and different L1/L2 overhead

Table 9.1.8-1 Comparison of FTP file download performance with different L1/L2 overhead

TTI	Additional L1/L2 overhead	FTP File size 100 KB		FTP File Size 500KB		FTP File Size 1MB	
		File Download Time (ms)	UPT (Mbps)	File Download Time (ms)	UPT (Mbps)	File Download Time (ms)	UPT (Mbps)
1ms	Baseline 0%	148	5.4	308	13.0	503	15.9
0.5ms	0%	79	10.1	234	17.1	434	18.4
	10%	81	9.9	254	15.7	474	16.9
	20%	86	9.3	281	14.2	529	15.1
	30%	99	8.1	331	12.1	609	13.1
0.1ms	40%	104	7.7	369	10.8	699	11.4
	0%	46	17.4	208	19.2	410	19.5
	10%	50	16.0	230	17.4	455	17.6
	20%	56	14.3	260	15.4	515	15.5
	30%	64	12.5	299	13.4	592	13.5
	40%	73	11.0	347	11.5	692	11.6

9.1.9 Simulation 9

Effect of UE and eNB processing times on TCP performance [6]

In current LTE, UE processing time for downlink data reception is determined based on PUCCH transmission timing. For example, in FDD, the UE should send HARQ-ACK in subframe n+4 where the UE receives PDSCH in subframe n. Therefore, the maximum UE processing time is considered to be 3ms (which is denoted by Δ_{UE}). This value is same as UE processing time for uplink as the UE sends uplink data at subframe n+4 when the UE receives Uplink grant in subframe n.

Note that the actual processing time available to the UE may be reduced due to TA (timing advance) (i.e., twice the propagation delay) as shown in Figure 9.1.9-1 below.

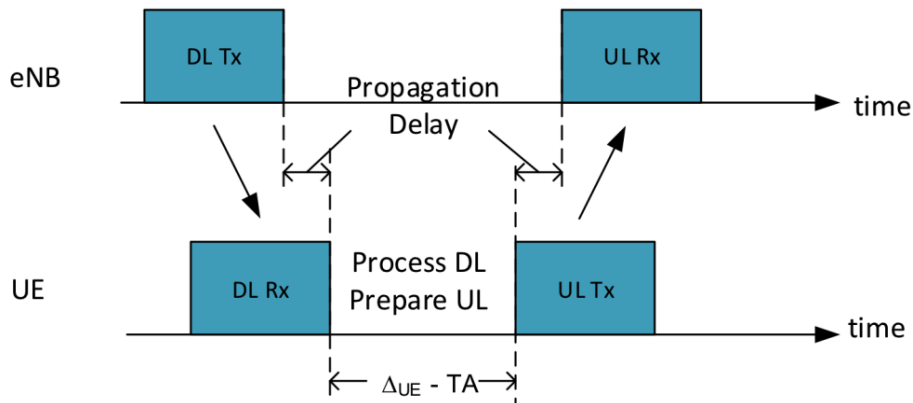


Figure 9.1.9-1 UE's actual available processing time is reduced due to propagation delays

In the following studies, we evaluate latency reduction gain with respect to different processing time in UE and in eNB side because the UP delay can be achieved by reducing UE/eNB processing time even in the normal TTI (1ms).

We assume decreased processing times of 2ms and 1ms (i.e., eNB takes 2ms / 1ms to process SR and send grant; UE takes 2ms / 1ms from receiving DL data to sending ACK).

One way to achieve less processing delay would be TBS (Transport Block Size) restriction by the specification in order to accelerate decoding speed together with potential decoder improvement by UE implementation. However, the exact solution can be discussed in RAN1.

9.1.9.1 Simulation assumptions

For this study, we consider 1ms and 0.5ms TTI lengths with 5TTI SR periodicity.

In this evaluation it is assumed that the core network and the Internet delay is very small and is independent of the TTI values and eNB/UE processing times. The assumed latency between eNB and TCP Server is 1-2 ms.

The average DL data-rate is kept the same for different scenarios. This is achieved by scaling the TBS sizes according to the duration of TTI compared to the current 1ms TTI.

For the FTP based file download using TCP protocol, TCP New Reno is used. File size is kept so as to achieve the saturation in data-rate so that the conditions when UE's perceived instantaneous as well as overall data-rate reaches at least 90% of the saturation rate can be determined.

The instantaneous throughput is calculated as 25ms moving window average of TCP layer throughput. Aggregate user perceived throughput (UPT) is calculated as total TCP layer data downloaded divided by the time elapsed. Initial time for setting up the network and for the UE to attach to network is removed as offset from the results such that the FTP request is sent to the server from the UE at time 0.

A saturation data-rate of approximately 21Mbps (example scenario is 3 CA 20MHz each, cell spectral efficiency ~3.5bps/Hz, 10 UEs/cell) is assumed. MCS index of UE is set such that on an average this data-rate is achieved when TCP has fully ramped up (channel adaptation is disabled).

The results for TTI reduction in this simulation do not take into account possible overhead due to e.g. additional L1/2 overhead.

9.1.9.1 Evaluation results

Figure 9.1.9-1 and Figure 9.1.9-2 show the comparison of TCP slow-start for different node processing delays. For simplicity, the processing delays for UE and eNB (denoted by Δ_{UE} and Δ_{eNB} respectively) are assumed to be equal in the following results. The performance of current processing delay (3ms) is compared with reduced values (2ms and 1ms) for UE and eNB processing delays. As expected the TCP ramps up quickly with smaller node processing times.

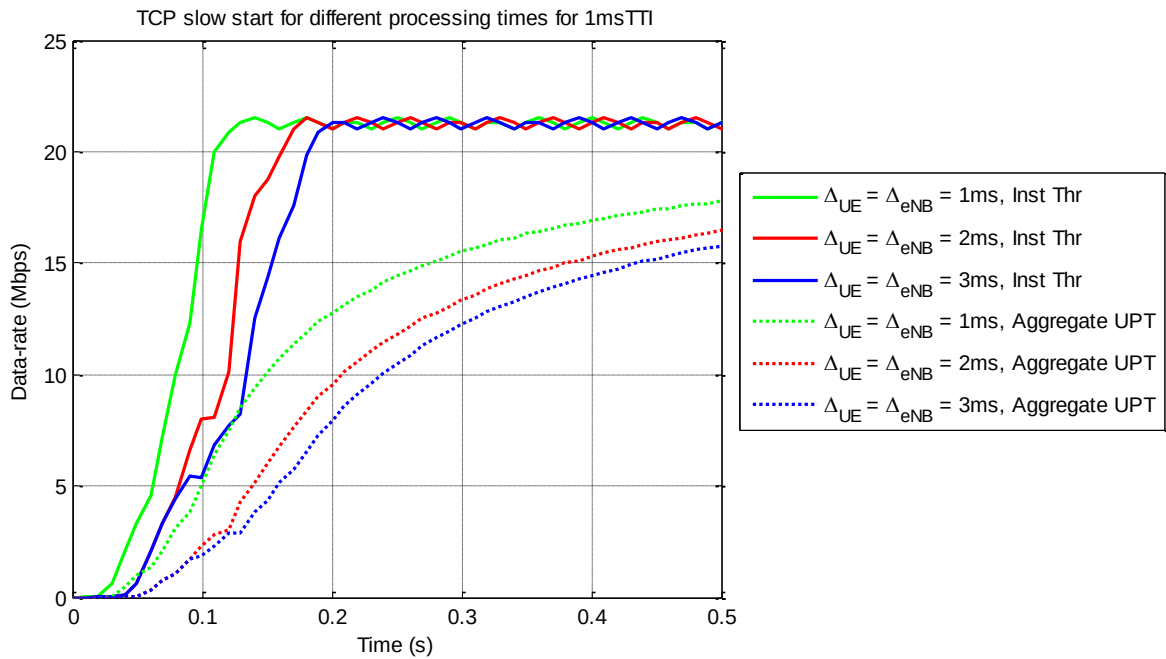


Figure 9.1.9-2 Effect of reduced node processing times on TCP slow-start behaviour with 1ms TTI

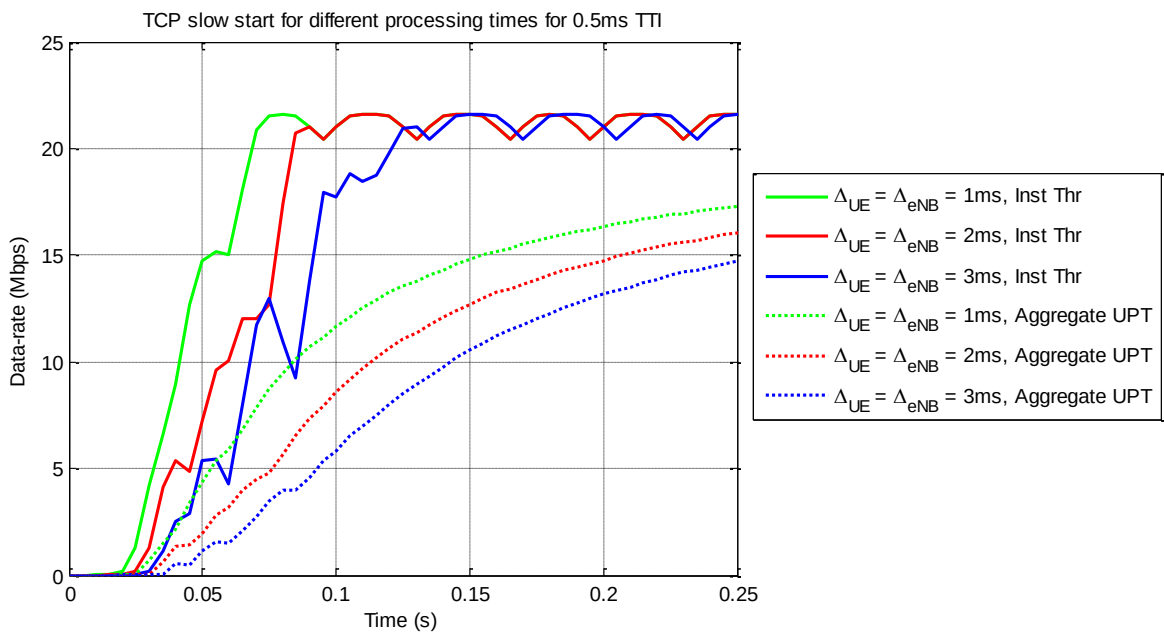


Figure 9.1.9-3 Effect of reduced node processing times on TCP slow-start behaviour with 0.5ms TTI

Table 9.1.9-1 shows the comparison of the scenarios described above. It is observed that about 40% gains are observed in the considered metrics when node processing delays are reduced to 1ms from baseline 3ms in 1ms TTI. Keeping the existing processing time (3ms from DL to UL), reduction in TTI from 1ms to 0.5ms reduces the TCP ramp-up time by 1/3rd. The ramp-up time can be decreased by more than half with 0.5ms TTI compared to 1ms TTI, even with a more relaxed 4TTI UE processing time.

Observation 33. TCP slow-start performance is improved with the reduced UE/eNB processing times.

Table 9.1.9-1 Comparison of TCP slow-start performance

$\Delta_{UE}, \Delta_{eNB}$ (ms)	Time for TCP ramp-up to 90% of saturation rate*	Min file size for TCP ramp-up to 90% of saturation rate	Time taken for UPT to be 90% of saturation rate*	Min file size for UPT to be 90% of saturation rate
1ms TTI				
3ms - baseline	176ms	142KB	1.31s	3.12MB
2ms	152ms (-14%)	124KB (-13%)	1.06s (-19%)	2.53MB (-19%)
1ms	107ms (-39%)	84KB (-41%)	768ms (-41%)	1.84MB (-41%)
0.5ms TTI				
3ms (6TTI)	116ms (-34%)	114KB (-20%)	813ms (-38%)	1.93MB (-38%)
2ms (4TTI)	82ms (-53%)	65KB (-54%)	644ms (-51%)	1.53MB (-51%)
1ms (2TTI)	67ms (-62%)	63KB (-56%)	457ms (-65%)	1.1MB (-65%)

It is also observed from Table 9.1.9-1 that the performance gain can be achieved by reducing processing times or by reducing TTI or both. In the studied scenarios, gains are observed to be comparable for keeping the 1ms TTI and reducing the processing time to 1ms (1TTI) or reducing the TTI to 0.5ms and keeping the same 3ms of processing time (i.e., 6TTI). On the other hand, as shown in earlier results, the gains in TCP ramp-up with reduced TTI of 0.5ms while keeping 3TTI processing time is significantly high (~60%) compared to 1ms TTI and 3TTI processing time. So, in general, TTI reduction shows more gain than simply reducing processing time without changing TTI.

Observation 34. TTI reduction is more effective than simply reducing processing time without TTI reduction.

It should be noted that with the change in processing time, the HARQ RTT changes (as shown in Appendix A1.4). Therefore different number of HARQ processes can be supported based on the TTI, as shown in Table 9.1.9-2 below.

Table 9.1.9-2 Number of possible HARQ processes

$\Delta_{UE}, \Delta_{eNB}$ (ms)	1ms TTI		0.5ms TTI	
	HARQ RTT (ms)	Max num of UL HARQ processes	HARQ RTT (ms)	Max num of UL HARQ processes
3ms	8	8	7	14
2ms	6	6	5	10
1ms	4	4	3	6

Table 9.1.9-3 below shows the average air-interface latency for UEs in connected mode. Average downlink latency and latency for the UE initiated uplink data transmission for a UE with uplink synchronization are calculated following the timing analysis of as explained in Appendix A1.4. As expected, both DL and UL average delays (SR-related steps and Data transmission steps) can be reduced with reduction in node processing times.

Table 9.1.9-3 Average Air-interface latency for different processing delays and TTI

$\Delta_{UE}, \Delta_{eNB}$ (ms)	Average air-interface latency							
	DL		UL (SR periodicity 5ms)					
	error free	10% error	error free			10% error		
			SR-related	Data Tx	Total	SR-related	Data Tx	Total
1ms TTI								
3ms	4ms	4.8ms	9ms	4ms	13ms	9ms	4.8ms	13.8ms
2ms	3ms	3.6ms	7.5ms	3ms	10.5ms	7.5ms	3.6ms	11.1ms
1ms	2ms	2.4ms	6ms	2ms	8ms	6ms	2.4ms	8.4ms
0.5ms TTI								
3ms	3.5ms	4.2ms	6.75ms	3.5ms	10.25ms	6.75ms	4.2ms	10.95ms
2ms	2.5ms	3ms	5.25ms	2.5ms	7.75ms	5.25ms	3ms	8.25ms
1ms	1.5ms	1.8ms	3.75ms	1.5ms	5.25ms	3.75ms	1.8ms	5.55ms

It is observed from the latency values in Table 9.1.9-3 and TCP ramp-up performance results in Table 9.1.9-1 that the TCP ramp-up performance is highly correlated, as expected, to the average air-interface latency, i.e., the ramp-up is faster if the air-interface latency is lower. In addition, the same benefits in terms of TCP slow-start performance can be observed with latency reduction in general regardless of how the lower latency is achieved.

9.1.10 Simulation 10

System Performance Gain with TTI reduction [6]

9.1.10.1 Faster UE Feedback and Rate Control

The current LTE TTI is 1ms and composed of two 0.5ms slots with seven symbols (in case of normal CP length) per slot. In selection of a new shorter TTI, keeping the necessary work, backwards compatibility and co-existence in mind, it should be chosen as a multiple of the symbol length. Therefore, here we assume that the shortest TTI possible is one symbol and use that to compare with the current TTI duration.

For the possible shorter TTIs, we will assume that processing and feedback times are scaled with TTI duration; for example the HARQ feedback is always after four TTIs. This is based on the RAN2#90 agreement that “*We assume that the UE and eNB processing time (HARQ RTT) scales with the TTI duration*”.

A shorter TTI can provide faster CSI and HARQ, which can lead to more accurate rate control algorithms that better match selected data rates with rapidly changing radio conditions. This is beneficial even for low mobility UEs since radio conditions can change quickly due to bursty interference from neighbour cells. As shown in Figure 9.1.10-1, the reported CSI and adjusted value after rate control is much closer to the actual channel conditions observed by the UE.

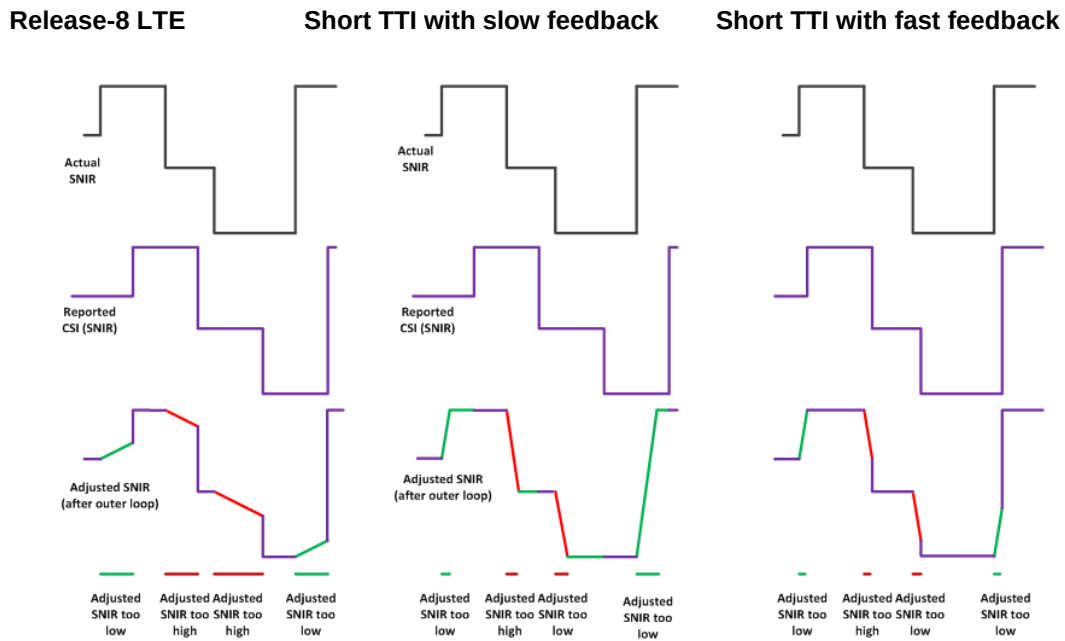


Figure 9.1.10-1 Benefits of Fast Rate Adaptation (slow feedback is not scaled with TTI while fast feedback is scaled with TTI)

9.1.10.2 Simulation Assumptions:

- 3GPP D1 Configuration:

	Freq (MHz)	Cell Radius (km)	Ant Height (Hb) (m)	Avg Clutter Height (m)	Dhb (m)	Slope	I	Avg EIRP (dBW) in 5 MHz	eNB Tx Pwr (dBW)	UE Ant Loss (dB)	Impl. Loss (dB)	Log normal Shadowing	Down tilt (deg)	Noise Figure (dB)	Penetration Loss (dB)	Receive Height (m)	Vert Beam width (deg)	Horiz Beam width (deg)
D1	2000	0.288	30	15	15	37.6	128.1	33	13	6	3	8	10	6	20	1.5	10	70

- 1 channel of 10Mhz with one LTE macro site of 3 sectors.
- 10 UEs are dropped randomly in each cell, 3 cells per site, 7 sites.
- 3 km/h slow fading.
- FTP: Poisson arrivals with
 - Case 1: inter arrival time = 0.1/sec, PktSize=100kbits
 - Case 2: inter arrival time = 1/sec, PktSize=1Mkbits
 - Case 3: inter arrival time = 4/sec, PktSize=4Mbits
- TTI durations: 1 symbol and 1ms
- HARQ ACK/NAK Delay = 4*TTI
- CQI Feedback Cycle = 4*TTI or 4ms (for TTI=1symbol, slow feedback assumes 4ms)

9.1.10.3 Simulation Results

System level simulation results under the 3GPP methodology for FTP traffic are shown in Table 9.1.10-1. The details of simulation assumptions are provided in the Appendix. TTI duration of 1 symbol was compared to the current 1ms TTI. The results are for the UE-perceived throughput and median delay, which are calculated as the average throughput, and delay for the received FTP packets. Three file sizes chosen for the FTP packets were 100Kbits, 1Mbits, and 4Mbits.

As seen in the results, even with the same absolute CSI feedback frequency, there is significant gain for the UE perceived throughput as well as median delay since the rate control algorithm can adapt better with faster HARQ turnaround. Further gains are also seen when the CSI feedback cycle is reduced. The gains are higher for smaller FTP packet sizes. This is expected since the total transmission time increase with larger packet sizes, which allows rate adaptation for 1ms TTI to improve.

Table 9.1.10-1 UE-perceived throughput comparison between 1 symbol and 1ms TTI

File size (100 Kbits)	5-th %ile tput	median tput	95-th %ile delay	median delay
LTE	4.8 Mbps	8.87 Mbps	20.8 ms	11.3 ms
Short TTI, slow CSI	2.32x	2.27x	0.43x	0.44x
Short TTI, fast CSI	4.39x	3.63x	0.23x	0.28x
File size (1 Mbits)	5-th %ile tput	median tput	95-th %ile delay	median delay
LTE	12.4 Mbps	22.4 Mbps	80.4 ms	44.7 ms
Short TTI, slow CSI	1.45x	1.44x	0.69x	0.70x
Short TTI, fast CSI	1.98x	1.77x	0.50x	0.57x
File size (4 Mbits)	5-th %ile tput	median tput	95-th %ile delay	median delay
LTE	15.0 Mbps	28.0 Mps	267 ms	143 ms
Short TTI, slow CSI	1.15x	1.17x	0.87x	0.85x
Short TTI, fast CSI	1.46x	1.27x	0.69x	0.79x

9.1.11 Simulation 11

TTI reduction gain with additional L2 overhead [9]

9.1.11.1 Simulation assumptions

The details of the TBS (Transport Block Size) calculation assumptions for short TTI duration are given in Annex A1.6. The detailed simulation assumptions including the header bits for each layer are given in Annex A1.6. The performance of 1ms TTI duration is the baseline. The L1 loss rate for the 1ms TTI duration is assumed as 8.3%, which is calculated as given in Annex A1.6. Only small file size (etc. 0.5Mbyte) is evaluated.

9.1.11.2 Evaluation results

The header overhead rate is the total header bits (from the TCP layer to the MAC layer) divided by the total TBS(s) used for data transmission. The header overheads in the RLC/HARQ retransmission are not considered for the calculation.

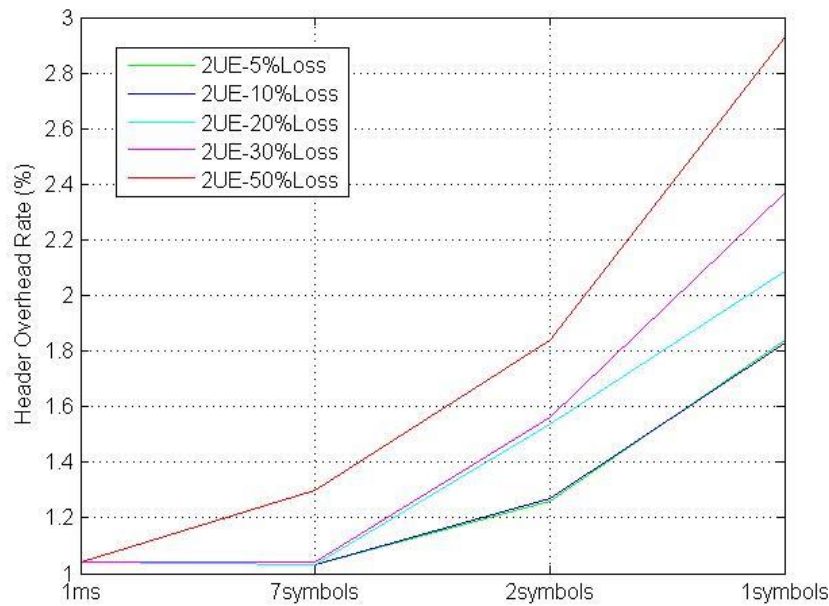


Figure 9.1.11-1: Header overhead rate

As the figure given above, the header overheads generally increase with the reduction of the TTI duration, as the shorter TTI duration leads to smaller TBS which in turn results in more segments in the RLC layer, and brings more header bits in RLC and MAC.

Observation 35. The header overheads increase with the reduction of TTI duration.

9.1.12 Simulation 12

TTI reduction gain with additional L2 overhead [10]

9.1.12.1 Simulation assumptions

In our analysis, TBS of legacy 14 OFDM symbol TTI length from 36.213 [8] is used as reference for calculating TBS for 7 symbols and 2 symbols TTI length, assuming same L1 overhead and same MCS.

Minimum L2 overhead is assumed with 1 byte PDCCP header, 1 byte RLC header, 1 byte MAC header and 3 byte CRC, i.e. in total 6 bytes for a TB.

Table 9.1.12-1 to 9.1.12-5 and Figure 9.1.12-1 to 9.1.12-5 illustrate the L2 overhead varies with TBS for TTI length of 14, 7 and 2 OFDM symbols, with 50 PRB and 5 PRB allocated.

9.1.12.2 Evaluation results

Table 9.1.12-1: L1 overhead with 50PRBs

<i>I</i> _{TBS}	TBS_14sym	TBS_7sym	TBS_2symb	L2 OH_14sym	L2 OH_7sym	L2 OH_2sym
0	1384	680	177	0.034	0.068	0.239
1	1800	888	237	0.026	0.053	0.184
2	2216	1096	296	0.021	0.043	0.150
3	2856	1416	387	0.017	0.033	0.117
4	3624	1800	497	0.013	0.026	0.092
5	4392	2184	607	0.011	0.022	0.076
6	5160	2568	717	0.009	0.019	0.065
7	6200	3088	865	0.008	0.015	0.054
8	6968	3472	975	0.007	0.014	0.048
9	7992	3984	1121	0.006	0.012	0.042
10	8760	4368	1231	0.005	0.011	0.038
11	9912	4944	1395	0.005	0.010	0.034
12	11448	5712	1615	0.004	0.008	0.029
13	12960	6468	1831	0.004	0.007	0.026
14	14112	7044	1995	0.003	0.007	0.024
15	15264	7620	2160	0.003	0.006	0.022
16	16416	8196	2325	0.003	0.006	0.020
17	18336	9156	2599	0.003	0.005	0.018
18	19848	9912	2815	0.002	0.005	0.017
19	21384	10680	3034	0.002	0.004	0.016
20	22920	11448	3254	0.002	0.004	0.015
21	25456	12716	3616	0.002	0.004	0.013
22	27376	13676	3890	0.002	0.004	0.012
23	28336	14156	4027	0.002	0.003	0.012
24	30576	15276	4347	0.002	0.003	0.011
25	31704	15840	4509	0.002	0.003	0.011
26	36696	18336	5222	0.001	0.003	0.009

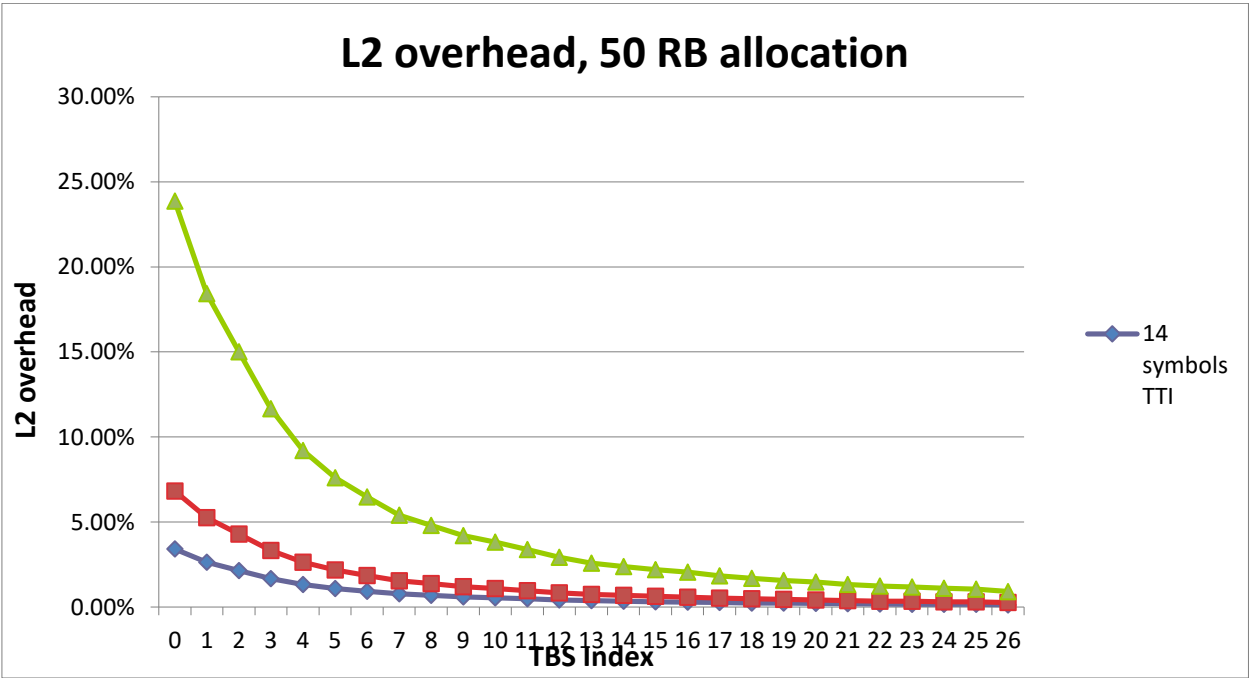


Figure 9.1.12-1: L2 overhead with 50PRB

Table 9.1.12-2: L1 overhead with 5PRBs

I_{TBS}	TBS_14sym	TBS_7sym	TBS_2sym	L2 OH_14sym	L2 OH_7sym	L2 OH_2sym
0	120	48	0	0.333	0.667	1.000
1	176	76	5	0.240	0.480	1.000
2	208	92	9	0.207	0.414	1.000
3	256	116	16	0.171	0.343	1.000
4	328	152	26	0.136	0.273	0.955
5	424	200	40	0.107	0.214	0.750
6	504	240	51	0.091	0.182	0.636
7	584	280	63	0.079	0.158	0.553
8	680	328	77	0.068	0.136	0.477
9	776	376	90	0.060	0.120	0.420
10	872	424	104	0.054	0.107	0.375
11	1000	488	122	0.047	0.094	0.328
12	1128	552	141	0.042	0.083	0.292
13	1256	616	159	0.038	0.075	0.263
14	1416	696	182	0.033	0.067	0.233
15	1544	760	200	0.031	0.061	0.214
16	1608	792	209	0.029	0.059	0.206
17	1800	888	237	0.026	0.053	0.184
18	1992	984	264	0.024	0.048	0.167
19	2152	1064	287	0.022	0.044	0.154
20	2344	1160	314	0.020	0.041	0.142
21	2472	1224	333	0.019	0.038	0.135
22	2664	1320	360	0.018	0.036	0.125
23	2856	1416	387	0.017	0.033	0.117
24	2984	1480	406	0.016	0.032	0.112
25	3112	1544	424	0.015	0.031	0.107
26	3752	1864	515	0.013	0.025	0.089

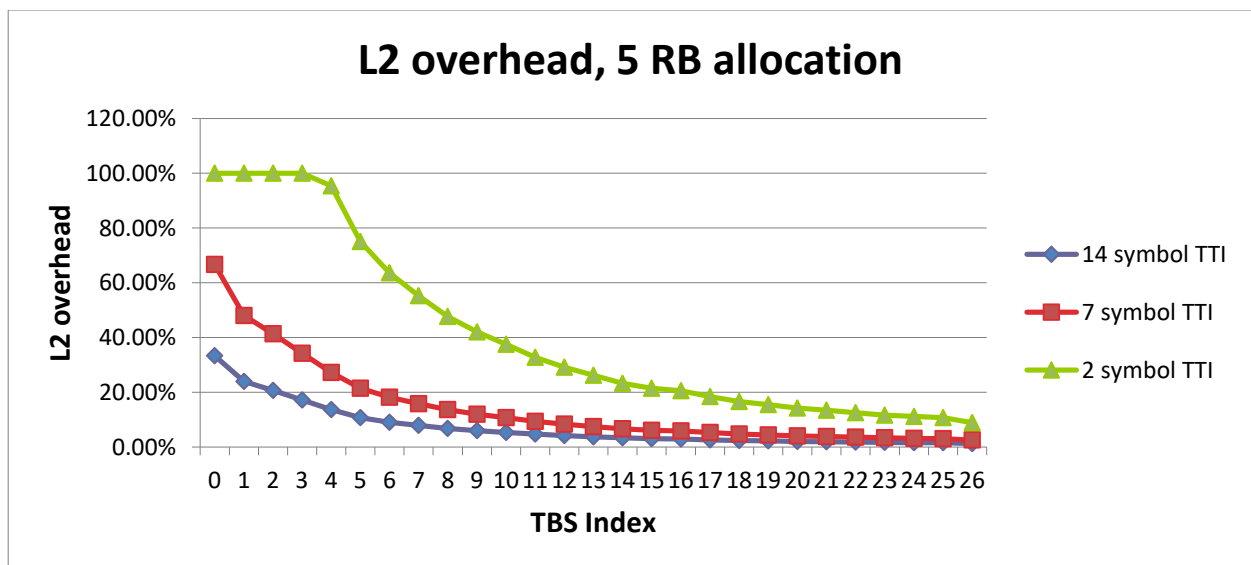


Figure 9.1.12-2: L2 overhead with 5PRB

Observation 36. L2 overhead increases significantly for shorter TTI length with smaller TBS.

Observation 37. L2 overhead is very high when the number of PRBs used for shorter TTI is limited and when low MCS is used, for some cases the TBS is not even enough just for the L2 headers and CRC.

9.2 Protocol evaluations on Contention based PUSCH transmission

9.2.1 Evaluation 1 on solution 1 [16]

In the following evaluations, we assume that 4 different UEs share the same PUSCH resource.

9.2.1.1 Resource efficiency

Assuming the UE packet arrival rate is 'a', the collision probability from system perspective can be calculated as follow:

$$6*a^2*(1-a)^2 + 4*a^3*(1-a) + a^4 \text{ (i.e. probability for 2, 3 and 4 colliding UEs respectively)}$$

Figure 9.2.1.1-1 illustrates the resource efficiency gain provided by contention based PUSCH transmission (compared to the pre-scheduling scheme allowed by current specifications for which UL grant is assigned every 1ms) and the corresponding collision probability with different UE packet arrival rates (i.e. 10% means in average one UL packet every 10ms). It is observed that the collision probability increases with the increase of UE packet arrival rate, and the resource efficiency gain decreases with the increase of collision probability. The radio efficiency gain is significant when the collision probability is low (e.g.<30%).

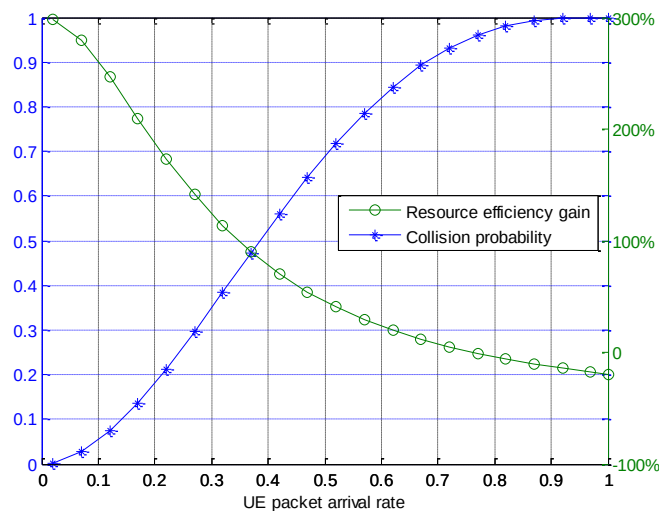


Figure 9.2.1.1-1: Resource efficiency gain and collision probability with different UE packet arrival rates

Note: Even if collision happens, it is still possible for the eNB to successfully decode some or all of the PUSCH transmissions, if the colliding UEs have uncorrelated radio channels. In this case, no extra PUSCH resource is needed for the retransmission.

Observation 38. Compared to the pre-scheduling scheme allowed by current specifications, contention based PUSCH transmission can significantly improve the radio efficiency when the collision probability is low (e.g.<30%).

9.2.1.2 Uplink latency

Table 9.2.1.2-1 provides the latency analysis on contention based PUSCH transmission if no collision happens, assuming that the contention based PUSCH resource is provided by dynamic scheduling every 1ms. The analysis shows that the uplink data transmission can be achieved within 8.5ms. In case the contention based PUSCH resource is provided by SPS, the uplink latency will be shorter (e.g. 6.5ms) as the UE doesn't need to receive and decode the UL grant on PDCCH.

Table 9.2.1.2-1: Uplink latency for contention based PUSCH transmission without collision (error free)

Component	Description	Time [ms]
1	Average delay to TTI border	0.5
2	eNB transmits UL Grant on PDCCH	1
3	UE Processing Delay (decoding of scheduling grant + L1 encoding of UL data)	3
4	Transmission of UL data	1
5	Data decoding and processing in eNB	3
	Total delay	8.5

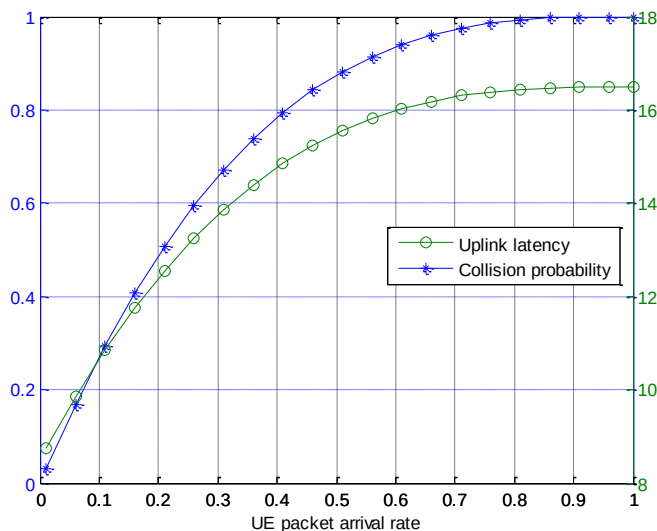
If collision happens for contention based PUSCH transmission, the colliding UEs will suffer additional delay caused by the retransmission over dedicated PUSCH resource. The total uplink latency can be calculated as follow:

$$T_{latency,CB} = 8.5ms + 8ms * P_{collision}, \text{ where } P_{collision} \text{ is the collision probability.}$$

Assuming the UE packet arrival rate is 'a', the collision probability $P_{collision}$ from UE perspective can be calculated as follow:

$$1 - (1-a)^3$$

Figure 9.2.1.1-2 illustrates the uplink latency of contention based PUSCH transmission and the corresponding collision probability with different UE packet arrival rates (i.e. 10% means in average one UL packet every 10ms). It is observed that the collision probability increases with the increase of UE packet arrival rate, and the uplink latency increases with the increase of collision probability. In case of low collision probability (e.g. 10%), the increased uplink latency is marginal and the total uplink latency is only 9.3ms. If the contention based PUSCH resource is provided by SPS, the total uplink latency will be even shorter (e.g. 7.3ms) in this case.

**Figure 9.2.1.1-2: Uplink latency and collision probability with different UE packet arrival rates**

Note: Even in case collision happens, it is still possible for the eNB to successfully decode some or all of the PUSCH transmissions, if the colliding UEs have uncorrelated radio channels. In this case, there is no additional uplink latency.

Observation 39. For contention based PUSCH transmission, the increased uplink latency due to collision is marginal if the collision probability is low (e.g. 0.8ms if the collision probability is 10%) and the maximum value is 8ms, assuming PUSCH resources for adaptive HARQ retransmission are available for all colliding UEs.

9.2.2 Evaluation on solution 2 [17]

9.2.2.1 Resource efficiency analysis on current solutions

Currently there are three uplink access solutions: D-SR (1ms/10ms D-SR period), SPS (1ms/10ms interval) and pre-scheduling. For D-SR solution, the uplink latency is given in Table 5.2.1-1 in section 5.2, and for SPS and pre-scheduling solutions, the uplink latency is given in Table 9.2.2.2-1 and Table 9.2.2.2-2.

Table 9.2.2.2-1: Typical uplink access latency component in case of SPS

Component	Description	Time (ms)
1	Average waiting time for configured UL grant (10 ms/1ms SPS interval)	5/0.5
2	UE Processing Delay (L1 encoding of UL data)	3
3	Transmission of UL data	1
4	Data decoding and processing in eNodeB	3
	Total delay [ms]	12/7.5

Table 9.2.2.2-2: Minimum uplink access latency component in case of pre-scheduling

Component	Description	Time (ms)
1	TTI alignment	0.5
2	UE Processing Delay (decoding of scheduling grant + L1 encoding of UL data)	3
3	Transmission of UL data	1
4	Data decoding and processing in eNodeB	3
	Total delay [ms]	7.5
Note: In pre-scheduling, eNB schedules UE to perform UL transmission blindly. When and how to perform blind schedule is totally up to eNB implementation. To achieve the minimum uplink latency, it is assumed that there is no extra delay between acquiring the uplink grant from eNB and obtaining data from upper layer, and eNB is required to give the uplink grant to the UE every TTI.		

To compare the resource efficiency and latency amongst the three solutions, same UE UL data transmission frequency is assumed. With the assumption of UE UL packet arrival rate 5% (i.e. average 1Tx/20ms), the comparison is given in Table 9.2.2.2-3.

Table 9.2.2.2-3: Comparison of the three current uplink access solutions

	D-SR	SPS	Pre-scheduling
Min Average UL latency	17ms/12.5ms (☹) (D-SR period 10ms/1ms)	12 ms/7.5ms (☹/☹) (SPS interval 10ms/1ms)	7.5 ms (☹) (pre-scheduling every 1ms)
Resource Utilization in 20ms	D-SR PUCCH cfg: 2/20 D-SR PUCCH Tx: 1 PDCCH: 1 PUSCH: 1 PHICH: 1	PUSCH: 2/20 PHICH: 2/20	PDCCH: 20 PUSCH: 20 PHICH: 20
Resource efficiency	High (☺)	Medium (☹)/ Low (☹)	Low (☹)

From Table 9.2.2.2-3, the observation can be given as below:

Observation 40. 1ms interval SPS and pre-scheduling have lowest uplink latency, but have worst resource efficiency.

Hence, based on the shorten SPS interval and pre-scheduling solution, some enhancements should be considered to improve resource efficiency, and the direct way is to share the resource with more than one UEs (i.e. CB-PUSCH).

9.2.2.2 Uplink latency

Similar with 1ms interval SPS, to achieve the lowest uplink latency, it is assumed that eNB allocates UL resource for one CB UE group every TTI. The uplink latency of CB PUSCH is given in Table 9.2.2.2-4.

Table 9.2.2.2-4: Uplink access latency component in case of CB PUSCH

Component	Description	Time (ms)
1	TTI alignment	0.5
2	UE Processing Delay (L1 encoding of UL data)	3
3	Transmission of UL data	1
4	Data decoding and processing in eNodeB	3
	Total delay [ms] (if no collision)	7.5
If collision	No feedback received	4
	Re-transmission with random backoff time	Random BO(<=10)
	Transmission of UL data	1
	Data decoding and processing in eNodeB	3
	Total delay [ms]	To be evaluated

Since the collision is due to more than one UE with uplink data transmission demand in the same time, two factors impact the collision probability: the UE number in one CB group and the transmitted packet arrival rate (PAR). Table 9.2.2.2-5 gives the collision probability based on different PAR and CB UE number.

Table 9.2.2.2-5: Collision probability for different PAR and CB UE number

	PAR=2%	PAR=5%	PAR=10%	PAR=15%	PAR=20%
2 CB UE	2.18%	6.05%	12.81%	21.36%	94.19%
3 CB UE	4.72%	12.29%	36.79%	99.70%	99.94%
4 CB UE	6.86%	19.37%	99.62%	99.94%	99.97%
5 CB UE	9.61%	27.36%	99.93%	99.97%	99.99%

Based on Table 9.2.2.2-4 and Table 9.2.2.2-5, the UL latency in different collision probability is given in Table 9.2.2.2-6.

Table 9.2.2.2-6: CB PUSCH latency with various collision probabilities

Collision Prob	Avrg UL Latency (ms)
2%	7.561
4%	7.798
6%	8.007
8%	8.283
10%	8.380
12%	8.589
14%	8.727
16%	8.854
18%	9.063
20%	9.186
22%	9.431
24%	9.482
26%	9.752
28%	10.043
30%	10.431
32%	11.558
34%	11.691
36%	11.737
38%	12.614
40%	13.053

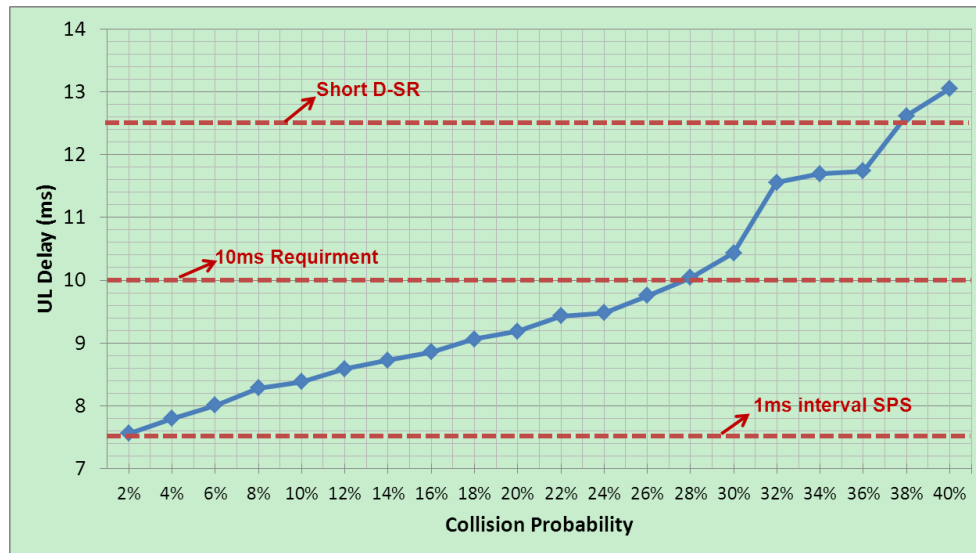


Figure-9.2.2.2-1 UL Latency performance with various collision probability

From Figure 9.2.2.2-1, it can be observed:

Observation 41. The higher collision probability, the higher UL latency.

Observation 42. Latency performance of CB PUSCH (every 1ms) is between D-SR (1ms period) and 1ms interval SPS if collision probability is low.

9.2.2.3 Resource efficiency

To analyse the resource efficiency and compare with shorten SPS interval, similar UL latency should be assumed for SPS and CB-PUSCH. The method in detail is given as below:

- 1 In case of different SPS interval, obtain the related UL latency and resource efficiency based on the calculation as Table 9.2.2.2-1;

Case	1	2	3	4	5
SPS Interval	1ms	2ms	3ms	4ms	10ms
Latency	7.5ms	8ms	8.5ms	9ms	12ms

- 2 To meet the similar latency, calculate the resource efficiency of CB PUSCH;

- a) Obtain the collision probability according to Figure-1;

Case	1	2	3	4	5
Latency	7.5ms	8ms	8.5ms	9ms	12ms
CB collision probability	2%	6%	10%	19%	37%

- b) Obtain the PAR and the CB UE number under the collision probability according to Table-5;

	2 CB UE	3 CB UE	4 CB UE	5 CB UE
PAR=2%	2.18% Case 1	4.72%	6.86% Case 2	9.61% Case 3
PAR=5%	6.05%	12.29%	19.37% Case 4	27.36%
PAR=10%	12.81%	36.79% Case 5	99.62%	99.93%

- c) Obtain the resource efficiency;

resource efficiency= (valid number of data transmission) / (total PUSCH resource for CB or SPS).

Note: valid number of data transmission = observation time * PAR5% * CB UE number;

The comparison results are given in Table 9.2.2.2-7 (observation time = 100ms) .

Table 9.2.2.2-7: Resource efficiency comparison

Case		Item	CB PUSCH	SPS
1	1ms SPS interval	Average UL latency	≈7.5 ms (Collision prob ≈ 2%)	7.5 ms
		Grant interval	1ms	1ms
		CB UE number	2	1
		Valid number of data transmission	4 (PAR=2%)	2
		Total PUSCH resource	100	100
		PUSCH efficiency	4%	2%
2	2ms SPS interval		CB PUSCH	SPS
		Average UL latency	≈8 ms (Collision prob ≈ 6%)	8 ms
		Grant interval	1ms	2ms
		CB UE number	4	1
		Valid number of data transmission	8 (PAR=2%)	2
		Total PUSCH resource	100	50
		PUSCH efficiency	8%	4%
3	3ms SPS interval		CB PUSCH	SPS
		Average UL latency	≈8.5 ms (Collision prob ≈ 10%)	8.5 ms
		Grant interval	1ms	3ms
		CB UE number	5	1
		Valid number of data transmission	10 (PAR=2%)	2
		Total PUSCH resource	100	33
		PUSCH efficiency	10%	6%
4	4ms SPS interval		CB PUSCH	SPS
		Average UL latency	≈9 ms (Collision prob ≈ 19%)	9 ms
		Grant interval	1ms	4ms
		CB UE number	4	1
		Valid number of data transmission	20 (PAR=5%)	5
		Total PUSCH resource	100	25
		PUSCH efficiency	20%	20%
5	10ms SPS interval		CB PUSCH	SPS
		Average UL latency	≈12 ms (Collision prob ≈ 37%)	12 ms
		Grant interval	1ms	10ms
		CB UE number	3	1
		Valid number of data transmission	30 (PAR=10%)	10
		Total PUSCH resource	100	10
		PUSCH efficiency	30%	100%

Analysis

From Table 9.2.2.2-7, it can be seen that the PUSCH efficiency improves with the increased UL latency, both for the CB PUSCH and SPS solution. With the low collision probability, CB PUSCH is better than SPS in PUSCH resource efficiency aspect, e.g. collision probability is lower than 10%. However, with higher collision probability, e.g. 37%, the PUSCH efficiency of CB PUSCH is worse than that of SPS. In summary, CB PUSCH solution can improve radio resource efficiency and provide good latency in case of lower collision probability.

Observation 43. Observation 4: With the same latency performance as 1ms interval SPS, CB PUSCH solution can improve radio resource efficiency in case of lower collision probability.

9.2.3 Evaluation 2 on solution 1 [18]

In the following section, we evaluate the collision probability and UL delay for CB-PUSCH transmissions. There are two cases: a) evaluation of CB-PUSCH collision rate, in a scenario with CB-PUSCH users only (i.e. in the system there

are only the CB-PUSCH users), b) evaluation of CB-PUSCH collision rate, while having in the system both CB-PUSCH only users and dynamically scheduled users.

Note: In both cases, the re-transmissions are dynamically scheduled on non-CB resources.

A) Evaluation of CB-PUSCH collision rate, in a scenario with only CB-PUSCH users

In this case, there is only one contention based group, to which the users could belong to in case they are using the CB-PUSCH resources. There are up to 6 UEs that could belong to this group, this assumption being based on the DMRS orthogonality achieved by allocating different cyclic shifts to UEs that can be supported by Rel-8.

In the evaluations, we have used UL traffic; with FTP model 3 [4] with static number of UE and packet arrival according to Poisson process, having the file size of 100 B. More parameters are given in Table A1.7-1 and Table A1.7-2 in Annex A1.7. The evaluation is done for different packet arrival rates (PAR), varying from low PAR 2% up to PAR 30% (i.e. approx. every 3rd TTI there is a file download).

In Figure 9.2.3.1 is shown the collision rate of CB-PUSCH and in Figure 9.2.3.2 is shown the average UL delay. The collision probability refers to collision probability of the transmissions.

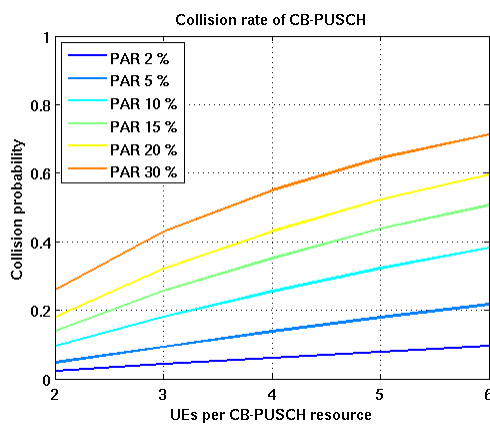


Figure 9.2.3.1 Collision rate of CB-PUSCH

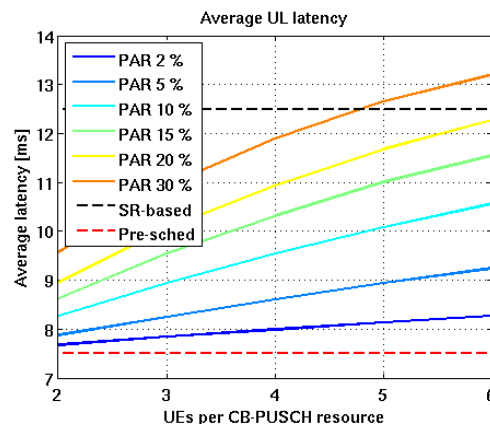


Figure 9.2.3.2 Average UL delay

From the results, we can observe that:

Observation 44.

- Only for PAR 2% the collision rate stays below 10%. Increasing the PAR, the collision rate increases up to ~70% in the case of PAR 30%. Of course, having higher number of users per group increases the probability of collisions.
- From the average UL latency point of view, taking as reference the legacy D-SR UL grant scheme with an average delay of 12.5 ms, we can get better performance in most of the cases i.e. unless there are very many collisions than the legacy D-SR UL grant scheme with an average delay of 12.5 ms, but it always has longer latency than 1ms pre-scheduling/SPS due to the collision.

The trade-off is UL overhead due to CB-PUSCH resources that are reserved even though not always used. Figure 9.2.3.3 shows the trade-off between the UL overhead and latency. The UL overhead here is relative to dynamic scheduling of all transmissions.

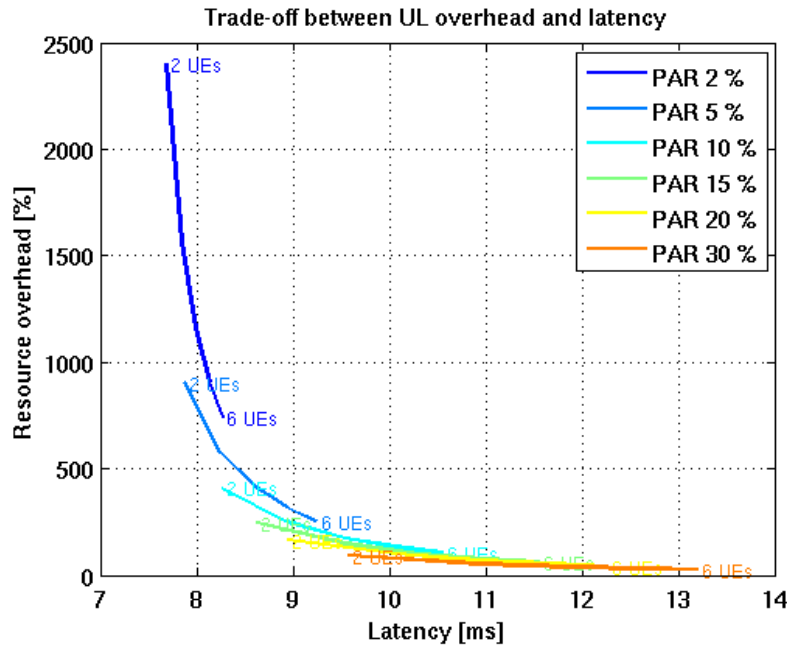


Figure 9.2.3.3 Trade-off between UL overhead and avg. latency

From the results we can observe that:

Observation 45.

- With low packet arrival rate, the CB-PUSCH resources are mostly unused, which means large resource overhead and inefficiency.
- With higher packet arrival rate the resource usage is more efficient, while the latency reduction benefit diminishes.

B) Evaluation of CB-PUSCH collision rate, in a scenario with CB-PUSCH and dynamically scheduled users

In this case, there are up to four contention-based groups, to which the users could belong to in case they are using the CB-PUSCH resources. Up to 4 UEs could belong to one group.

We have used FTP model 1 [3], with UE arrival according to Poisson process, having the file size of 0.5 MB. More parameters are given in Table A1.7-1 and Table A1.7-3 in Annex A1.7.

In the evaluations, we used DL traffic, having in UL TCP ACKs. This represents as the ‘best performance’ case traffic for the CB-PUSCH, as increasing the UL traffic the collisions and the delay would only increase even more.

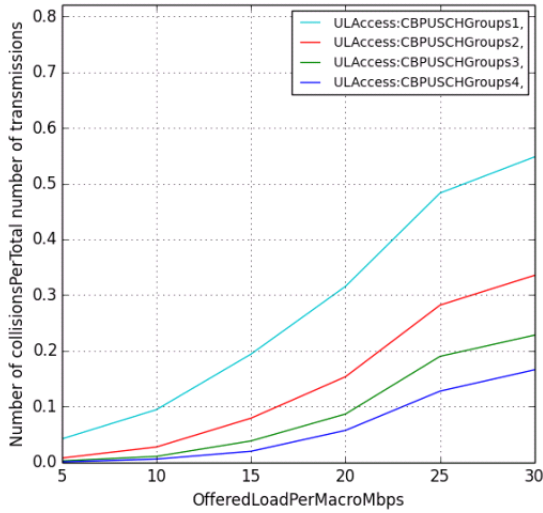


Figure 9.2.3.4 Collision rate

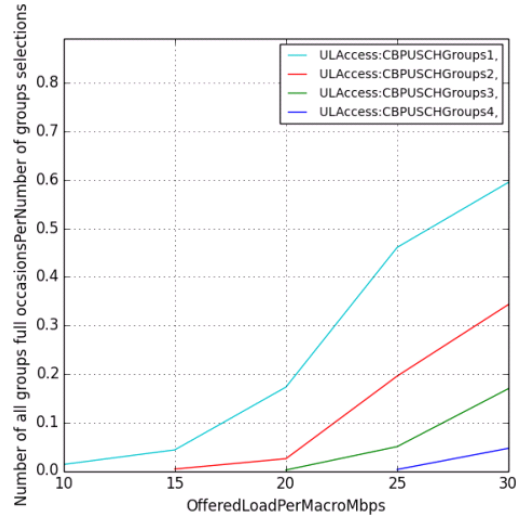


Figure 9.2.3.5 Probability of full CB PUSCH groups

From the results we can observe that:

Observation 46.

- There is a high impact of the number of CB-groups per TTI: the lower the number of CB-groups, the higher the collision probability (e.g. in case of one CB-group, there is a 60% probability of collision, while with 4 CB-groups the collision probability could reach up to 18%)
- The collision probability depends on the load in the NW (simultaneously active UE/cell) and also number of RB reserved for the CB
- From the probability that all available CB PUSCH groups are full for different loads and numbers of groups:
 - With 5 Mbps of offered DL load the probability of filling even one group of 4 UEs is zero
 - Two groups can tolerate up to 15 Mbps offered load without getting full, three groups up to 20 Mbps and four groups up to 25 Mbps

One key point to observe it is the UL delay (E2E delay). In that case, our reference case is legacy D-SR UL grant (with SR periodicity 5 ms). This aspect is plotted for selected cases, corresponding for 10 Mbps and 25 Mbps offered load.

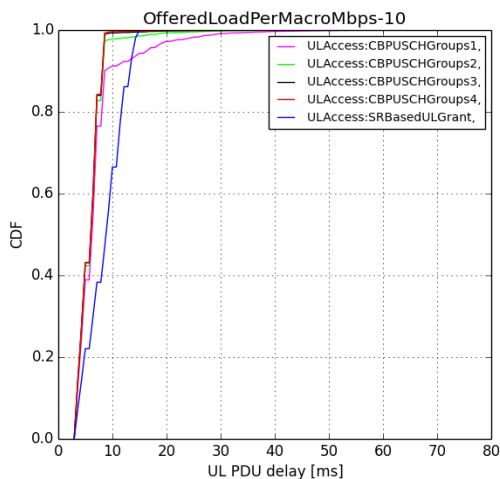


Figure 9.2.3.6 UL end to end delay for 10 Mbps

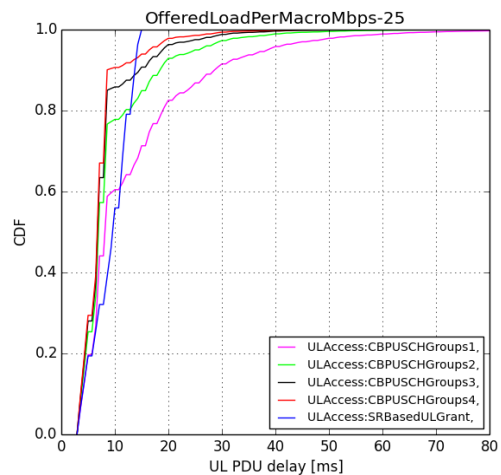


Figure 9.2.3.7 UL end to end delay for 25 Mbps

From the results we can observe that:

Observation 47. SR-based UL grant gives more predictable delay performance whereas the gains from CB-PUSCH are sensitive to load and number of UEs sharing the same resources. High end of delay distribution is significantly worse for CB-PUSCH, especially in high load.

9.3 Handover latency [11]

Figure 9.3-1 captures typical interruption time for legacy handover vs. gains achieved by Synchronous handover. As can be seen from the figure 9.3-1 below, the interruption time is much higher for UE's with more speed and the case of higher number of small cells under a Macro cell. This is due to the fact that if the UE moves with higher speed, it will experience more frequent handovers than a UE with lower speed. The increased number of handovers, each introducing some interruption time adds up to a higher overall delay. However, the interruption time is significantly lower for the case of synchronized handover. The fact remains, that the higher number of handover a UE experiences, the higher will be the data interruption time. However, comparing the two scenarios, the benefits are obvious with synchronized handover.

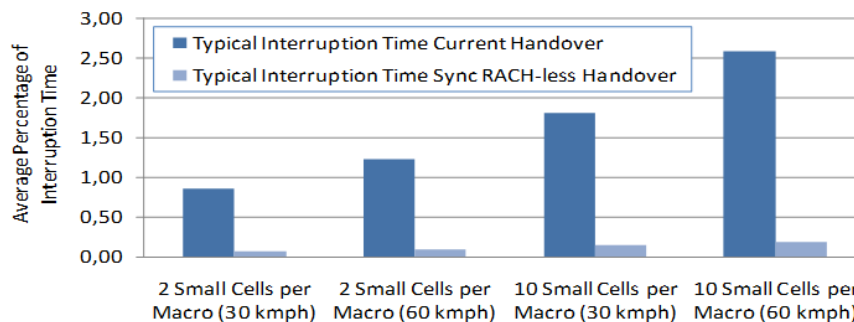


Figure 9.3-1: Average Percentage of the Interruption Time.

9.4 Findings from system evaluations on TTI reduction and reduced processing time

9.5 Findings from link evaluations on TTI reduction and reduced processing time

10 Conclusion

10.1 RAN2 Protocol Evaluations

In protocol evaluations, the potential gains in terms of increased download throughput and reduction of download time for reduced latency in LTE has been performed and evaluated.

For TCP, and TCP slow start specifically, results show that reduced UL latency, shorter RTT and HARQ RTT can have a positive impact on TCP performance depending on cell load and L1/L2 overhead. This is due to the fact that the receiver may acknowledge TCP packets faster which enables a faster increase in the TCP window size. Latency reduction has shown a positive impact on both TCP congestion avoidance mode and in TCP slow start phase.

As the initial window size for each TCP connection is very small and the increase is steeper for each size increment, the effect of latency reductions for both RTT and HARQ RTT are more considerable for the slow start phase. This impact is

large for small file sizes, especially where the slow start period lasts for the entire duration of the file transfer (how long the phase lasts depends on TCP congestion control algorithm and parameters). For larger file sizes, the proportion of the slow start phase of the whole file is smaller if packet losses can be avoided.

Furthermore, the following high-level observations have been obtained:

For TTI shortening, the potential gain depends on the following aspects:

- Delays in the core network and Internet also impact the UE perceived throughput. If these delays dominate over radio network delays, latency reduction techniques will be less efficient.
- In high loaded cells, or in situations where the available scheduled Uu bitrate is low, e.g. due to radio conditions, user data is queued before transmitting. This has an impact on User Throughput. In these scenarios, efforts to reduce radio latency will be less important as the limited availability of UL resources may result in reduced performance as the queuing delay dominates over the radio delay.
- By reducing the TTI length, the network can schedule the UE faster, which reduces the RTT. A reduction in RTT increases the TCP throughput. A reduction of TTI length may also increase the system capacity for small data transmission.
- For file transfers using TCP, the amount of UPT gain provided by shorter TTI may decrease with larger file size.
- With reduced TTI length, the processing time may be reduced with a resulting scaling of the RTT and HARQ RTT. Shortening the TTI gives a larger latency reduction effect compared to only shortened processing time(s).
- A reduction of TTI length may require additional L1 overhead e.g. CRC, RS, DCI, UCI and result in a relative increase of L2 headers. Increased overhead will to some extent limit the User Throughput gain from TTI reduction.
- Shorter TTI can enable quicker HARQ and CSI feedback, which may give faster link adaptation to channel conditions. Gain can be observed for the UE perceived throughput as well as median delay.

For fast uplink access:

- An enhancement to SPS to allow for periodic UL grants every TTI (Fast UL) reduces the latency of the first UL transmission compared to legacy intervals, and performs equally well compared with SR every 1ms with lower control channel load. Fast UL improves User Throughput also with shorter TTI, although the relative gain is smaller compared to longer TTIs.
- An enhancement of UL grants allowing the UE to skip padding transmissions in grant if it has no UL data in the UE buffer is beneficial as it may decrease UL interference and improve UE battery efficiency.
- The gains of acknowledged activation/deactivation of SPS has not been established and needs to be discussed further in a work item phase.

SR based PUSCH transmission and the pre-scheduling scheme allowed by current specifications could give more predictable delay performance whereas the delay of contention based PUSCH transmission are sensitive to the collision probability, which is subject to the number of UEs sharing the same PUSCH resource (e.g. depends on the cell load) and the UE's traffic arrival interval. Contention based PUSCH transmission allows more efficient PUSCH resource utilization compared to the pre-scheduling scheme, but with increased uplink latency if collision occurs. When the collision probability is low, the increase in uplink latency is small. In this case the PUSCH resource utilization is lower than that of the SR based PUSCH transmission because the contention based PUSCH resources reserved for the UEs may not be used.

It would be beneficial if the NB can identify the UE that performed the PUSCH transmission even when collision happens, e.g. by different DMRS Cyclic Shift. RAN2 evaluation for contention based PUSCH transmission assumes that the eNB can always detect the DMRS resources whenever there are transmissions from UEs, but the actual performance would depend on success rate of DMRS resource detection, which should be studied by RAN1.

For handover latency reduction:

- The following steps contribute to a major portion of total handover delay and can be addressed for possible latency reduction:
 - RACH procedure including delay to acquire first available PRACH in target cell, PRACH preamble transmission and UL allocation + TA,

- UE processing time after RA procedure including decoding of scheduling grant and timing alignment + L1 encoding of UL data, and transmission of *RRCConnectionReconfigurationComplete*.
- Two potential solution directions which may be beneficial were identified during the study. The feasibility of these solutions has not been studied by other WGs. No further work is expected in this study.

Annex A: Simulation assumptions

Simulation assumptions for the latency reduction study are provided within this Annex. The simulation assumption is divided in three parts, i.e. protocol simulations assumptions, system simulation assumptions and link-level simulation assumptions.

A1 Protocol Simulations

For each source for protocol evaluations the corresponding assumptions are provided within this section.

A1.1 Simulation 1 [4]

Average down-link latency calculation

Following the same approach as in section B.2.1 in 3GPP TR 36.912, the LTE U-plane one-way latency for a scheduled UE consists of the fixed node processing delays and 1 TTI duration for transmission, as shown in Figure A.1 below. Assuming the processing times can be scaled by the same factor of TTI reduction keeping the same number of HARQ processes, the one way latency can be calculated as

$D = 1.5 \text{ TTI (eNB processing and scheduling)} + 1 \text{ TTI (transmission)} + 1.5 \text{ TTI (UE processing)} + n \cdot 8 \text{ TTI (HARQ retransmissions)}$

$$= (4 + n \cdot 8) \text{ TTI.}$$

Considering a typical case where there would be 0 or 1 retransmission, and assuming error probability of the first transmission to be p , the delay is given by

$$D = (4 + p \cdot 8) \text{ TTI.}$$

So, for 0% BLER, $D = 4 \cdot \text{TTI}$,

And for 10% BLER, $D = 4.8 \cdot \text{TTI}$.

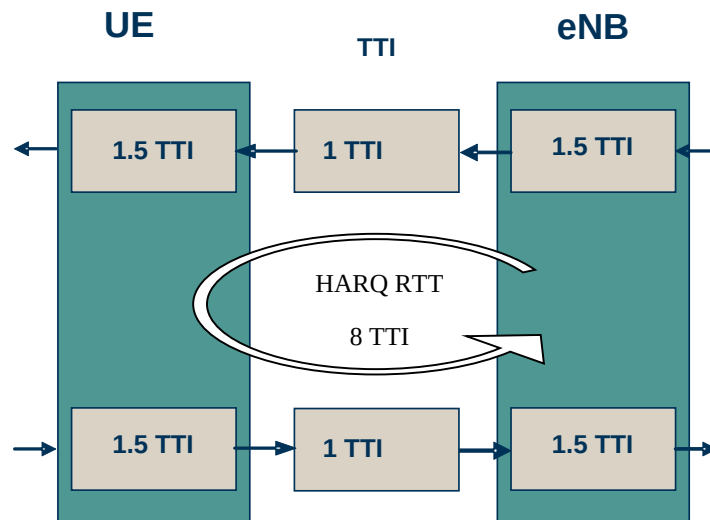


Figure A.1 eNB and UE processing delays and HARQ RTT

Average UE initiated UL transmission latency calculation

Assume UE is in connected/synchronized mode and wants to do UL transmission, e.g., to send TCP ACK. Following table shows the steps and their corresponding contribution to the UL transmission latency. To be consistent in comparison of DL and UL, we add the eNB processing delay in the UL after the UL data is received by the eNB (step 7).

Table A.1 UL transmission latency calculation

Step	Description	Delay
1.	Average delay to next SR opportunity	SR periodicity/2
2.	UE sends SR	1 TTI
3.	eNB decodes SR and generates scheduling grant	3 TTI
4.	Transmission of scheduling grant (assumed always error free)	1 TTI
5.	UE processing delay (decoding Scheduling grant + L1 encoding of data)	3 TTI
6.	UE sends UL transmission	$(1 + p \cdot 8)$ TTI where p is initial BLER.
7.	eNB receives and decodes the UL data	1.5 TTI

In the table above, steps 1-4 and half delay of step 5 is assumed to be due to SR, and rest is assumed for UL data transmission in values shown in Table 4

A1.2 Simulation 2 [5]

Table A.2-1: Simulation assumptions on radio aspects

Parameter	Assumption				
Simulation time	51s(Ftp1), 100s(HTTP)				
System bandwidth	20 MHz				
Duplex mode	FDD				
Carrier frequency	2GHz				
Cell layout	Hexagonal grid, 7 sites, 3 cells per site, with wrap-around				
Number of UEs	1, 5, 10, 20 per cell				
Inter-site distance	500m				
UE speed	3Km/h				
Antenna configuration	2 tx , 4 rx (CELL) ; 1 tx, 2 rx (UE)				
Antenna pattern (horizontal) (For 3-sector cell sites with fixed antenna patterns)	$A_H(\varphi) = -\min \left[12 \left(\frac{\varphi}{\varphi_{3dB}} \right)^2, A_m \right]$ $\varphi_{3dB} = 70 \text{ degrees}, A_m = 25 \text{ dB}$				
Antenna pattern (vertical) (For 3-sector cell sites with fixed antenna patterns)	$A_V(\theta) = -\min \left[12 \left(\frac{\theta - \theta_{etilt}}{\theta_{3dB}} \right)^2, SLA_v \right]$ $\theta_{3dB} = 10, SLA_v = 20 \text{ dB}$ The parameter θ_{etilt} is the electrical antenna downtilt. The value for this parameter, as well as for a potential additional mechanical tilt, is not specified here, but may be set to fit other RRM techniques used. For calibration purposes, the values $\theta_{etilt} = 15$ degrees for 3GPP case 1 and $\theta_{etilt} = 6$ degrees for 3GPP case 3 may be used. Antenna height at the base station is set to 32m. Antenna height at the UE is set to 1.5m.				
Combining method in 3D antenna pattern	$A(\varphi, \theta) = -\min \{ -[A_H(\varphi) + A_V(\theta)], A_m \}$				
Total BS TX power (Ptotal)	46 dBm – 20MHz carrier				
UE power class	23dBm (200mW) This corresponds to the sum of PA powers in multiple Tx antenna case				
Channel model	Large Fading channel model				
Distance-dependent pathloss	$L = 128.1 + 37.6 \log_{10}(R)$ (R in km)				
Lognormal shadowing model	Reference to B1.4.1.4 in UMTS TR30.03				
Lognormal shadowing standard deviation	8dB				
Correlation distance of shadowing	50 ms				
Shadowing correlation	<table border="1"> <tr> <td>Between sites</td><td>0.5</td></tr> <tr> <td>Between cells</td><td>1</td></tr> </table>	Between sites	0.5	Between cells	1
Between sites	0.5				
Between cells	1				
Penetration loss	20dB				
CQI measurement period	5 ms				
SRS reporting period	5 ms				
Number of RLC ARQ max transmit	16				
Number of MAC HARQ max transmit	3				
TP layer	TCP				
CN Delay	0ms (assuming eNB and APP server are co-located)				
BSR Period	5ms				
SR Period	1ms				

Table A.3-1: TCP parameters [3]

Parameter	Assumption
TCP ACK Delay	0ms

Parameter	Assumption
Initial TCP Window	1460 Bytes
Initial Ssthresh	65535 Bytes
Ssthresh	65536 Bytes
TCP Connection Number	1(FTP1), 5(HTTP)

Table A.4-1: FTP traffic model [3]

Parameter	Assumption
FTP Model	Model 1
Packet Size	0.5Mbytes
User arrival rate λ	Poisson distributed with arrival rate λ , $\lambda = 0.2$

Table A.5-1: HTTP traffic model [4]

Component	Distribution	Parameters	PDF
Main object size (SM)	Truncated Lognormal	Mean = 10710 Bytes SD = 25032 Bytes Min = 100 Bytes Max = 2 MBytes	$f_x = \frac{1}{\sqrt{2\pi}\delta x} \exp \left[-\frac{(\ln x - \mu)^2}{2\delta^2} \right], x \geq 0,$ $\delta = 1.37, \mu = 8.37$ if $x > \max$ or $x < \min$, discard and generate a new value for x .
Embedded object size(SE)	Truncated Lognormal	Mean = 7758 Bytes SD = 126188 Bytes Min = 50 Bytes Max = 2 Mbytes	$f_x = \frac{1}{\sqrt{2\pi}\delta x} \exp \left[-\frac{(\ln x - \mu)^2}{2\delta^2} \right], x \geq 0,$ $\delta = 2.36, \mu = 6.17$ if $x > \max$ or $x < \min$, discard and generate a new value for x .
Number of embedded objects per page (Nd)	Truncated Pareto	Mean = 5.64 Max = 53	$f_x = \frac{\alpha_k^\alpha}{x^{\alpha+1}}, k \leq x < m$ $f_x = \left(\frac{k}{m}\right)^\alpha, x = m$ $\alpha = 1.1, k = 2, m = 55$ Subtract k from the generated random value to obtain N_d . If $x > \max$, discard and regenerate a new value for x .
Reading time	Exponential	Mean = 30 sec	$f_x = \lambda e^{-\lambda x}, x \geq 0$ $\lambda = 0.033$
Parsing time	Exponential	Mean = 0.13sec	$f_x = \lambda e^{-\lambda x}, x \geq 0$ $\lambda = 7.69$

A1.3 Simulation 4 [6]

For one symbol length TTI, we assume new PDCCH will be used for data scheduling. Although the legacy PDCCH is not needed anymore, we keep one symbol length legacy PDCCH region for CSS.

As we evaluated in system simulation for 3GPP Case 1, the average PDCCH Aggregation Level is 1.42. Therefore, we assume that the average PDCCH payload for each UE is 52 REs.

Table A1.3-1: L1 overhead analysis (CRS based DL channel estimation)

TTI length	1ms	1 symbol
Number of legacy UEs (not capable of one symbol length TTI) scheduled per ms	20	0
Number of new UEs (capable of one symbol length TTI) scheduled per ms	0	20
Number of CRS port	2	
Number of DMRS port	0	
System Bandwidth (MHz)	20	
Average PDCCH payload for each UE (number of RE)	52	
CRS overhead in non-PDCCH region (number of RE/PRB/ms)	12	
Legacy PDCCH overhead for legacy UE scheduling (number of RE/ms)	1040	0
Total legacy PDCCH overhead (number of RE/ms)	2400 (2 symbols)	1200 (1 symbol)
New PDCCH overhead for new UE scheduling (number of RE/ms)	0	1040
Total PDCCH overhead (number of RE/ms)	2400	2240
CRS overhead (number of RE/ms)	1200	1200
Total overhead (number of RE/ms)	3600	3440
Overhead ratio	21.43%	20.48%
Normalized L1 data rate (compared to 1ms TTI)	100%	101.21%

Table A1.3-2: L1 overhead analysis (DM-RS based DL channel estimation)

TTI length	1ms	1 symbol
Number of legacy UEs (not capable of one symbol length TTI) scheduled per ms	20	0
Number of new UEs (capable of one symbol length TTI) scheduled per ms	0	20
Number of CRS port	1	
Number of DMRS port	2	
System Bandwidth (MHz)	20	
Average PDCCH payload for each UE (number of RE)	52	
CRS overhead in non-PDCCH region (number of RE/PRB/ms)	6	
DMRS overhead (number of RE/PRB/TTI)	12	4
Legacy PDCCH overhead for legacy UE scheduling (number of RE/ms)	1040	0
Total legacy PDCCH overhead (number of RE/ms)	2400 (2 symbols)	1200 (1 symbol)
New PDCCH overhead for new UE scheduling (number of RE/ms)	0	1040
Total PDCCH overhead (number of RE/ms)	2400	2240
CRS overhead (number of RE/ms)	600	600
DMRS overhead (number of RE/ms)	1200	5200
Total overhead (number of RE/ms)	4200	8040
Overhead ratio	25.00%	47.86%
Normalized L1 data rate (compared to 1ms TTI)	100%	69.52%

A1.4 Simulation 9 [6]

Downlink

It has been understood by RAN2 that processing delays in the different nodes form the major part of the delay in obtaining valid scheduling assignment, and reducing such processing delay could improve the latency compared to LTE Rel-8 [TR 36.912]. Following the same approach as in section B.2.1 in [TR 36.912], the LTE U-plane one-way latency for a scheduled UE consists of the fixed node processing delays and 1TTI duration for transmission, as shown in Figure B.1 below. Let us denote total processing time for eNB from UL reception to DL transmission (in ms) by Δ_{eNB} and total processing time for UE from DL reception to UL transmission (in ms) by Δ_{UE} . The fraction of these processing times needed for processing the reception and preparing the transmission are denoted by w , x , y and z where $w + x = 1$ and $y + z = 1$.

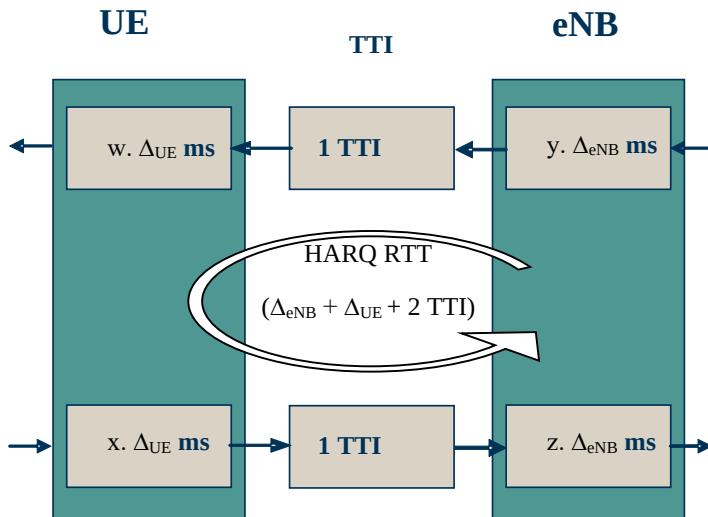


Figure A.1.4.1: eNB and UE processing delays and HARQ RTT

The one way latency can be calculated as

$D = y \cdot \Delta_{\text{eNB}} + 1 \text{ TTI} + w \cdot \Delta_{\text{UE}} + n * (\Delta_{\text{eNB}} + \Delta_{\text{UE}} + 2 \text{ TTI})$, where n is the number of HARQ retransmissions.

Considering a typical case where there would be 0 or 1 retransmission, assuming error probability of the first transmission to be p , and $w = x = y = z = 1/2$, the delay is given by

$$D = (1/2 + p) * (\Delta_{\text{eNB}} + \Delta_{\text{UE}} + 2 \text{ TTI}).$$

So, for 0% BLER, $D = 0.5 * (\Delta_{\text{eNB}} + \Delta_{\text{UE}} + 2 \text{ TTI})$, and for 10% BLER, $D = 0.6 * (\Delta_{\text{eNB}} + \Delta_{\text{UE}} + 2 \text{ TTI})$.

UE initiated UL transmission

Assume UE is in connected/synchronized mode and wants to do UL transmission, e.g., to send TCP ACK. Following table shows the steps and their corresponding contribution to the UL transmission latency.

Table A.1.4.1 UL transmission latency calculation

Step	Description	Delay
1.	Average delay to next SR opportunity	SR periodicity/2
2.	UE sends SR	1 TTI
3.	eNB decodes SR and generates scheduling grant	Δ_{eNB}
4.	Transmission of scheduling grant (assumed always error free)	1 TTI
5.	UE processing delay (decoding Scheduling grant + L1 encoding of data)	Δ_{UE}
6.	UE sends UL transmission	$1 \text{ TTI} + p * (\Delta_{\text{eNB}} + \Delta_{\text{UE}} + 2 \text{ TTI})$ where p is initial BLER.
7.	eNB receives and decodes the UL data	$z \cdot \Delta_{\text{eNB}}$

In the table above, steps 1-4 and half delay of step 5 is assumed to be due to SR, and rest is assumed for UL data transmission in values shown in Table 9.1.9-3.

A1.5 Simulation 5 [6]

Baseline L1 overhead calculation

We assume 20% L1 overhead for baseline case, which is calculated as following:

- TTI = 14 OFDM symbol
- Control overhead: 2 PDCCH OFDM symbols (24 REs)
- RS Overhead: 2 CRS antenna ports (additional 12 REs)

Total number of REs per PRB: $14 \times 12 = 168$

Overhead: $36/168 \approx 21\%$

TBS calculation

L1 overhead (i.e. L1 control channel and RS) is considered as 10%, 20%, 30% and 50%, which is reflected into the TBS calculation. To unify the TBS calculation, we also use the same TBS calculation for the base line case (i.e. 1ms TTI), and the 20% L1 overhead is assumed.

TBS is calculated by following formula based on efficiency in table below (R1-081638).

$$\text{TBS} = \text{PRB_Num} * \text{RE_per_PRB} * \text{efficiency} * \text{L1_Overhead}$$

Where,

- PRB_Num is available PRB number we set in simulation;
- RE_per_PRB is the total RE number in a PRB, e.g., for 14 OFDM Symbol TTI, RE_per_PRB is $12 \times 14 = 168$, and for 7 OFDM symbol TTI, RE_per_PRB is $12 \times 7 = 84$;
- Efficiency is for calculating the TBS table in TS 36.213, and efficiency = code rate * modulation level;
- L1_Overhead denotes the L1 signaling and RS overhead.

Note: CRC impact not considered.

MCS Index	modulation	coding rate x 1024	efficiency	Comments	Code Rate
0	2	120	0.2344	from CQI table	0.1171875
1	2	157	0.3057	Average Efficiency	0.15332031
2	2	193	0.377	from CQI table	0.18847656
3	2	251	0.4893	Average Efficiency	0.24511719
4	2	308	0.6016	from CQI table	0.30078125
5	2	379	0.7393	Average Efficiency	0.37011719
6	2	449	0.877	from CQI table	0.43847656
7	2	526	1.0264	Average Efficiency	0.51367188
8	2	602	1.1758	from CQI table	0.58789063
9	2	679	1.3262	Average Efficiency	0.66308594
10	4	340	1.3262	overlap	0.33203125
11	4	378	1.4766	from CQI table	0.36914063
12	4	434	1.69535	Average Efficiency	0.42382813
13	4	490	1.9141	from CQI table	0.47851563
14	4	553	2.1602	Average Efficiency	0.54003906
15	4	616	2.4063	from CQI table	0.6015625
16	4	658	2.5684	Average Efficiency	0.64257813
17	6	438	2.5684	overlap	0.42773438
18	6	466	2.7305	from CQI table	0.45507813
19	6	517	3.0264	Average Efficiency	0.50488281
20	6	567	3.3223	from CQI table	0.55371094
21	6	616	3.6123	Average Efficiency	0.6015625
22	6	666	3.9023	from CQI table	0.65039063
23	6	719	4.21285	Average Efficiency	0.70214844
24	6	772	4.5234	from CQI table	0.75390625
25	6	822	4.8193	Average Efficiency	0.80273438
26	6	873	5.1152	from CQI table	0.85253906
27	6	910	5.33495	Average Efficiency	0.88867188
28	6	948	5.5547	from CQI table	0.92578125

UL latency components

Normally, we assume that the UL access latency is stated in following tables:

Normal D-SR UL Latency component

Component	Description	Time
1	Average waiting time for PUCCH (5 TTI SR period)	2.5 * TTI
2	UE sends Scheduling Request (SR) on PUCCH	1 * TTI
3	eNB decodes Scheduling Request and generates the Scheduling Grant	3 * TTI
4	Transmission of Scheduling Grant	1 * TTI
5	UE Processing Delay (decoding of grant + L1 encoding of UL data)	3 * TTI
6	Transmission of UL data	1 * TTI
7	Data decoding and processing in eNodeB	3 * TTI
8	Backhaul(CN/Internet) transmission (to App server)	5ms
Total		14.5TTI+5ms

Short D-SR UL Latency component

Component	Description	Time
1	Average waiting time for PUCCH (1 TTI SR period)	0.5 * TTI
2	UE sends Scheduling Request (SR) on PUCCH	1 * TTI
3	eNB decodes Scheduling Request and generates the Scheduling Grant	3 * TTI
4	Transmission of Scheduling Grant	1 * TTI
5	UE Processing Delay (decoding of grant + L1 encoding of UL data)	3 * TTI
6	Transmission of UL data	1 * TTI
7	Data decoding and processing in eNodeB	3 * TTI
8	Backhaul(CN/Internet) transmission (to App server)	5ms
Total		12.5TTI+5ms

Pre-allocation UL Latency component

Component	Description	Time
1	Average waiting time for PDCCH	0.5 * TTI
2	UE reads Resource Blocks on PDCCH	1 * TTI
3	UE Processing Delay (decoding of grant + L1 encoding of UL data)	3 * TTI
4	Transmission of UL data	1 * TTI
5	Data decoding and processing in eNodeB	3 * TTI
6	Backhaul(CN/Internet) transmission (to App server)	5ms
Total		8.5TTI+5ms

A1.6 Simulation 11 [9]

The UL transmission latency and the HARQ RTT are based on the assumptions given in A1.1.

Table A1.6-1 Simulation assumptions

Parameter	Assumption
Simulation time	20s(Ftp model 2)
TTI duration	1, 2, 7 OFDM symbol; 1ms (legacy baseline)
System bandwidth	10 MHz
Duplex mode	FDD
Carrier frequency	2GHz
Cell layout	Hexagonal grid, 7 sites, 3 cells per site
UE distribution	Random distribution; average number of UE per cell: 2
Inter-site distance	500m
HARQ RTT	8TTI
L1 overhead	For 1, 2, 7 OFDM symbol: 5%, 10%, 20%, 30%, 50%
UE mobility	0 Km/h
Reserved OFDM symbol	0, 1 for legacy PDCCH
One-way CN Delay	1ms
Target BLER	10%
Downlink scheduler	Proportional fair in time and frequency
Antenna configuration	2 tx , 2 rx (CELL) ; 1 tx, 1 rx (UE)
Maximum Tx Power of eNodeB	46 dBm
UE power class	23dBm (200mW)
Channel model	Large Fading channel model
TCP parameters	Initial TCP window: 1460 MSS: 1460 Ssthresh: 65535

FTP traffic model	FTP Traffic Model 2 File size: 0.5 Mbytes Reading Time: Exponential Distribution, Mean = 1s $f_D = \lambda e^{-\lambda D}, D \geq 0$
Overheads of headers	TCP header: 20 bytes IP header: 20 bytes PDCP header: 2 bytes TCP+IP header overheads after RoHC: 8 bytes RLC header: 2 bytes MAC header: 2 bytes

TBS calculation assumption

The TBS (Transport Block Size) of each short TTI PDSCH is calculated as follows:

$$\mathbf{TBS}_{\text{short}} = \mathbf{TBS}_{\text{legacy}} \times \frac{\mathbf{N}_{\text{short}}}{\mathbf{N}_{\text{legacy}}} \times \frac{1 - \mathbf{L}_{\text{short}}}{1 - \mathbf{L}_{\text{legacy}}}$$

$\mathbf{TBS}_{\text{short}}$: The TBS of short TTI PDSCH

$\mathbf{TBS}_{\text{legacy}}$: The TBS of legacy PDSCH which uses the same MCS and PRB(s) as the short TTI PDSCH.

$\mathbf{N}_{\text{short}}$: The number of REs in short TTI PDSCH per PRB

$\mathbf{N}_{\text{legacy}}$: The number of REs in legacy PDSCH per PRB

$\mathbf{L}_{\text{short}}$: The loss rate of PHY layer within the short TTI PDSCH due to the REs not used for data transmission

$\mathbf{L}_{\text{legacy}}$: The loss rate of PHY layer within the legacy PDSCH due to the REs not used for data transmission

Resource mapping of short TTI

In Figure A1.6-1 the resource map above is the legacy resource mapping per PRB in one subframe, considering 2 Antenna ports and 2 OFDM symbols control field. In Figure A1.6-1 the resource map below is the short TTI resource mapping, considering 2 OFDM symbols used for the control field in order to ensure the backward compatibility. The loss rates ($\mathbf{L}_{\text{legacy}}$, e.g. 5% - 50%) of the PHY layer in short TTI duration are assumed.

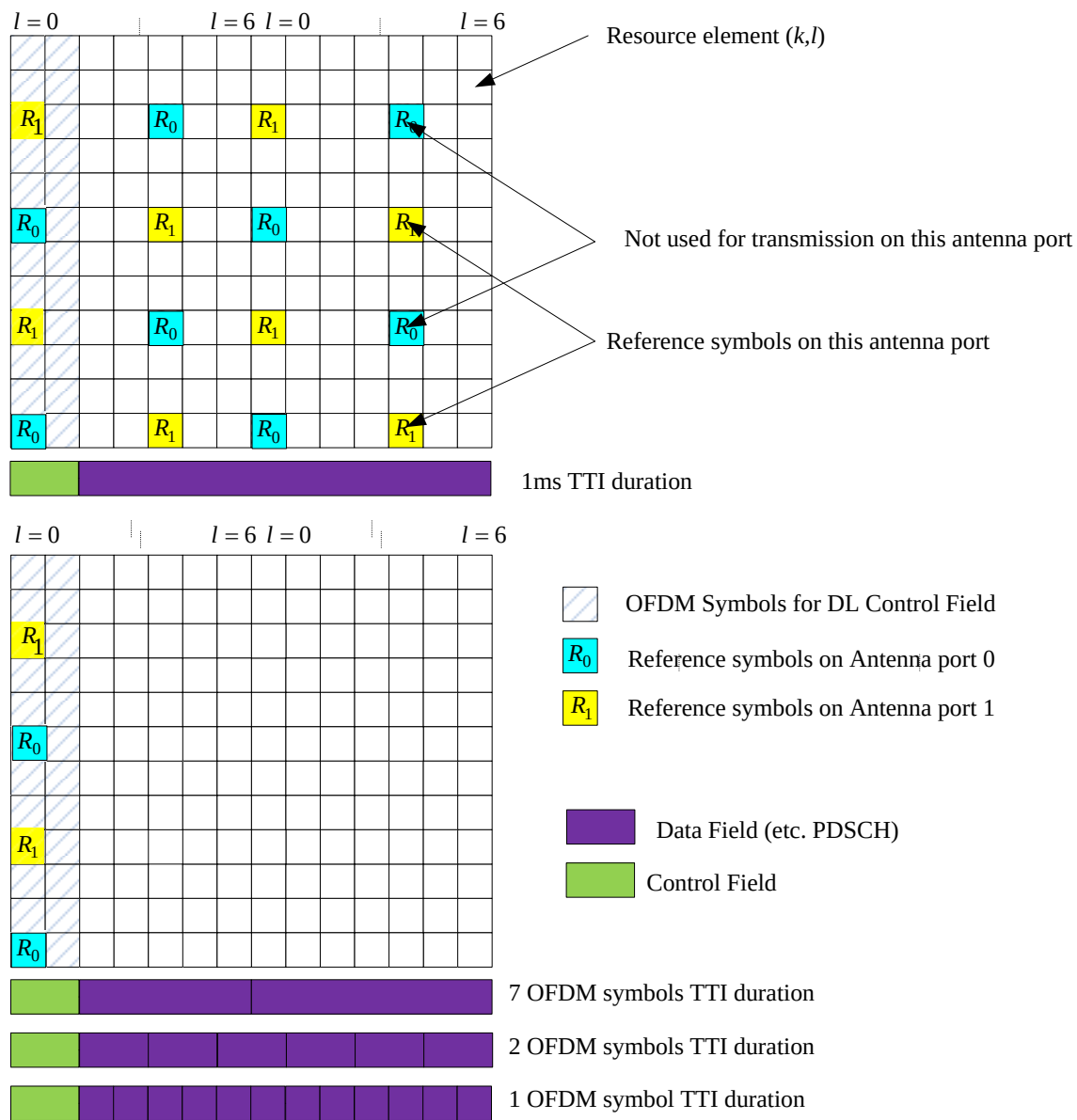


Figure A1.6-1: resource mapping per PRB in one subframe

TBS Calculation of short TTI

According to the resource mapping and the TBS calculation formula given above, the loss rate of PHY layer for legacy PDSCH is calculated as follows:

$$L_{\text{legacy}} = \frac{\text{the number of reference symbols within PDSCH}}{\text{the number of REs within PDSCH}} = \frac{12}{144} = 8.3\%$$

For different short TTI duration, The TBS of short TTI PDSCH is calculated as the following table:

Table A1.6-2: TBS calculation for different TTI duration

TTI Duration	TBS of short TTI PDSCH (TBS_{short})
7 OFDM symbol	First time slot: $TBS_{short} = TBS_{legacy} \times \frac{60}{144} \times \frac{1 - L_{short}}{1 - 8.3\%}$
	Second time slot: $TBS_{short} = TBS_{legacy} \times \frac{84}{144} \times \frac{1 - L_{short}}{1 - 8.3\%}$
2 OFDM symbol	$TBS_{short} = TBS_{legacy} \times \frac{24}{144} \times \frac{1 - L_{short}}{1 - 8.3\%}$
1 OFDM symbol	$TBS_{short} = TBS_{legacy} \times \frac{12}{144} \times \frac{1 - L_{short}}{1 - 8.3\%}$

A1.7 Simulation on contention based PUSCH transmission

The common evaluation assumptions are given in Table A1.7-1 below.

Table A1.7-1 Common evaluation assumptions

Parameter	Assumption
System bandwidth	20 MHz
Duplex mode	FDD
Carrier frequency	2GHz
Cell layout	Hexagonal grid, 7 sites, 21 cells per site, with wrap-around
Inter-site distance	500m
UE speed	3 km/h, quasi-static model
Antenna configuration	1x2 MRC
eNB TX power	46 dBm
UE Tx power	23dBm
Channel model	Typical Urban
Pathloss model	25.814
Lognormal shadowing, std. dev.	8dB
Penetration loss	20dB
HARQ RTT	8ms
SR Period	5 ms
DRX	OFF
Maximum number of scheduled users per TTI	10
L1 overhead	20%
TTI Length	14 symbols

The dedicated evaluation assumptions for scenario with only CB PUSCH users are given in Table A1.7-2 below.

Table A1.7-2 evaluation assumptions for only CB PUSCH users

Parameter	Assumption
Number of UEs	2, 3, 4, 5, 6 per cell (per CB PUSCH group)
FTP packet size	100 bytes
Packet arrival rate λ	FTP model 3 with packet arrival according to Poisson process
Core network delay	0ms
UL Load per macro cell	Packet arrival rate: 2, 5, 10, 15, 20, 30%
Delays: CB PUSCH model	Grant reception: 1ms UE processing delay: 3 ms eNB processing delay: 3ms Retransmission delay: 8 ms
CB PUSCH capacity	Number of groups: 1 Number of RBs per group: 10 Maximum number of UEs per group: 2, 3, 4, 5, 6

The dedicated evaluation assumptions for scenario with CB PUSCH and dynamically scheduled users are given in Table A1.7-3 below.

Table A1.7-3 evaluation assumptions for CB PUSCH and dynamically scheduled users

Parameter	Assumption
Number of UEs	According to offered load
SR to grant	4ms – scaled down with shorter TTI
Initial TCP Window	3 x 1500 Bytes (MSS), RFC 5681, section 3.1
Initial Ssthresh	45 x 1500 Bytes (MSS)
Ssthresh	Dynamic according to RFC 5681, sections 3.1 and 3.2
FTP file size	0.5 MB
User arrival rate λ	FTP model 1 with UE arrival according to Poisson process
Core network delay	2ms
DL Load per macro cell	5, 10, 15, 20, 25, 30 Mbps
UL Load per macro cell	TCP acks for the DL traffic above
Scheduler	TD: PF, FD: PF
Delays: SR based UL grant model	SR to grant reception: 4ms UE processing delay: 3 ms Average waiting time for SR: 2.5 ms
Delays: CB PUSCH model	Grant reception: 1ms UE processing delay: 3 ms Retransmission delay: 8 ms
CB PUSCH capacity	Number of groups: 1, 2, 3, 4 Number of RBs per group: 20 Maximum number of UEs per group: 4

A2 System simulation assumptions for reduced TTI and processing delay

Table A2-1: System simulation assumptions for reduced TTI and processing delay

Parameter	Assumptions
Layout	Both 7 and 19 Macro eNBs can be used, 3 sectors per site; Small cell scenario 2a as optional
System bandwidth per carrier	10MHz/20MHz
Carrier frequency	2GHz
Inter-site distance	500m
Total BS TX power (P _{total} per carrier)	46dBm
TTI length	1/2/3/4/7 symbols; other TTI lengths provided by companies Note that variable symbols and other numbers are not precluded Baseline: Fixed TTI length(s) across the legacy TTIs is assumed for 1 UE
Fast UL Access schemes	Optional; provided by companies
RS and control signaling overhead	Details of RS and control signaling overhead provided by companies
TBS determination	Scalable with TTI length as baseline
HARQ RTT	Scalable with TTI length as baseline; HARQ RTT not scaled with TTI provided by individual company
Scheduler	Proportional fairness
Distance-dependent path loss	ITU UMa[referring to Table B.1.2.1-1 in TR36.814], with 3D distance between an eNB and a UE
Penetration	For outdoor UEs:0dB
	For indoor UEs: 20dB+0.5din (din: independent uniform random value between [0, min(25,d)] for each link)
Shadowing	ITU UMa according to Table A.1-1 of 36.819 with 3D distance for shadowing correlation distance
Antenna pattern	3D, referring to TR36.819
Antenna Height:	25m
UE antenna Height	1.5m
Antenna gain + connector loss	17 dBi
Antenna gain of UE	0 dBi
Fast fading channel between eNB and UE	ITU UMa according to Table A.1-1 of 36.819
Antenna configuration	(mandatory) 2Tx(eNB), (optional) 8Tx(eNB), Cross-polarized 2Rx(UE), Cross-polarized
Number of UEs	10 UEs per macro cell Mixture of latency reduction capable UEs and legacy UEs is not precluded (Companies should provide details on how these UEs are handled in the simulations)
UE dropping	Randomly and uniformly dropped throughout the macro geographical area. 20% UEs are outdoor and 80% UEs are indoor.
Traffic model	FTP model 2 or 3 File size [100kbits, 100kB, 500kB, 1 MB] RU [20%, 40% 60%]
CSI report period	5, 10 TTIs and milliseconds between two consecutive reports Note: Companies should provide details of CSI measurement
CSI report delay	6 TTIs and milliseconds
TCP models	TCP Reno model (RFC 2581)

	- SSThresh 65535 Bytes - Initial window size 1460 Bytes - Max segment size 1460 Bytes 40 Bytes TCP header are added to the initial window size and max segment size The three way handshake is not modeled as baseline. TCP ACK feedback modeling is provided by the companies
UE receiver	MMSE-IRC; other UE receiver provided by companies
eNB noise figure	5dB
UL antenna configuration	(mandatory) 2Rx(eNB), (optional) 8Rx(eNB), 1Tx(UE)
UE noise figure	9dB
UE speed	3km/h, 60km/h, 120km/h (optional)
Duplex mode	FDD and TDD
Network synchronization	Synchronized
Core, transport and internet network delay	0ms, 6ms, 10ms
Performance metrics	Mean, 5%, 50% and 95% user perceived throughput Mean, 5%, 50% and 95% user packet delay

- User perceived throughput (UPT) is the average of all its file throughputs
- File throughput = file size/time needed to download the file
 - Time needed to download the file starts when the packet is generated, and ends when the last bit of the packet is correctly delivered to the receiver, the network delay (core, transport and internet network delay) is included here
- User packet delay is the average of all its file delays
 - File delay is the time needed to download the file as described above
- Unfinished files are not incorporated in the UPT and user packet delay calculation.
- MBSFN subframe configuration can be considered.

A3 Link level simulation assumptions

Table A3-1: PDSCH link-level simulation assumptions

Parameter	Value
Carrier frequency	2 GHz
System bandwidth	10 MHz/20MHz
TTI length	1/2/3/4/7 symbols; other TTI lengths provided by companies
Allocated bandwidth	25 PRBs, 50 PRBs
Channel model	EPA, EVA or ETU
UE speed	3km/h, 60km/h, 120km/h(optional)
Antenna configuration	2Tx(eNB), 2Rx(UE), 8Tx(eNB) optional
Antenna correlation	Uncorrelated
Legacy PDCCH region	2 OFDM symbols
CP length	Normal
Transmission mode	TM4, TM9
RS configuration	TM4: 2 CRS ports; TM9: 2 CRS ports, legacy DMRS pattern for 7 symbols; Detailed DMRS pattern provided by companies for 1/2/3/4 symbols
Receiver type	MMSE; other UE receiver provided by companies
Channel estimation	Practical
Rank adaptation	Fixed Rank
Link adaptation	Disabled
Modulation and code rate	64QAM 5/6, 16QAM 3/4, QPSK 1/3; others provided by companies
Precoding codebook	Fixed
TBS determination	Scale TBS size with TTI length
HARQ retransmission	Disabled (mandatory), Enabled (optional)
Performance metrics	BLER, MSE of channel estimation (optional)

Table A3-2: PUSCH link-level simulation assumptions

Parameter	Value
Carrier frequency	2 GHz
System bandwidth	10 MHz/20MHz
TTI length	1/2/3/4/7 symbols; other TTI lengths provided by companies
Allocated bandwidth	25 PRBs, 50 PRBs
Channel model	EPA, EVA or ETU
UE speed	3km/h, 60km/h, 120km/h(optional)
Antenna configuration	1Tx(UE), 2Rx(eNB), 8Rx(eNB) as optional
CP length	Normal
Transmission mode	TM1
DMRS configuration	Detailed DMRS pattern provided by companies
Receiver type	MMSE; other UE receiver provided by companies
Channel estimation	Practical
Link adaptation	Disabled
Modulation and code rate	64QAM 5/6, 16QAM 3/4, QPSK 1/3; others provided by companies
TBS determination	Scale TBS size with TTI length
HARQ retransmission	Disabled (mandatory), Enabled (optional)
Performance metrics	BLER, MSE of channel estimation (optional)
Fast UL Access schemes	Optional: provided by companies

Table A3-3: (E)PDCCH link-level simulation assumptions

Parameter	Value
Carrier frequency	2 GHz
System bandwidth	10 MHz/20MHz
TTI length	1/2/3/4/7 symbols; other TTI lengths provided by companies
Channel model	EPA, EVA or ETU
UE speed	3km/h, 60 km/h, 120km/h (optional)
Antenna configuration	2Tx(eNB), 2Rx(UE)
Payload size	40bits, 70bits as starting point including CRC; others provided by companies
CP length	Normal
Shortened (E)PDCCH design	Details provided by companies
Receiver type	MMSE; other UE receiver provided by companies
Channel estimation	Practical
Performance metrics	BLER, MSE of channel estimation (optional)

Table A3-1: PUCCH link-level simulation assumptions

Parameter	Value
Carrier frequency	2 GHz
System bandwidth	10 MHz/20MHz
TTI length	1/2/3/4/7 symbols; other TTI lengths provided by companies
Channel model	EPA, EVA or ETU
UE speed	3km/h, 60km/h, 120 km/h (optional)
Antenna configuration	1Tx(UE), 2Rx(eNB), 8Rx (eNB) (optional)
CP length	Normal
Shortened PUCCH design	Details provided by companies
Receiver type	MMSE; others provided by companies
Channel estimation	Practical
Performance metrics	For HARQ-ACK: ACK missed detection probability (1%), NACK-to-ACK error probability (0.1%); DTX-to-ACK probability 1%; For CQI: Probability of CQI block error $\leq 1\%$; MSE of channel estimation (optional)

Annex B: System evaluation results

Annex C: Link-level evaluation results

Annex <X>: Change history

This is the last annex for TRs which details the change history using the following table.

This table can be used for recording progress during the WG drafting process till TSG approval of this TR.

Change history							
Date	TSG #	TSG Doc.	CR	Rev	Subject/Comment	Old	New
2015-05	RAN2 #90	R2-152493			Draft skeleton TR	0.0.0	
2015-05	RAN2 #90	R2-152905			Editorial changes to structure	0.0.1	0.0.2
2015-05	RAN2 #90	R2-152918			Agreed TR Skeleton	0.0.2	0.1.0
2015-06	RAN2 #90	R2-152935			Agreed Updated TR as a result of email discussion [90#11][LTE/Latency]	0.1.0	0.1.1
2015-06	RAN2 #90	R2-152936			TR 36.881 v0.2.0 for Study on latency reduction techniques for LTE following result of email discussion [90#11][LTE/Latency]	0.1.1	0.2.0
2015-08	RAN2 #91	R2-153910			TR 36.881 v0.2.1 for Study on latency reduction techniques for LTE following result of email discussion [91#17][LTE/Latency]	0.2.0	0.2.1
2015-09	RAN2 #91	R2-153984			Agreed TR 36.881 v0.3.0 for study on latency reduction techniques for LTE as a result of email discussion [91#17][LTE/Latency]	0.2.1	0.3.0
2015-10	RAN2#91bis	R2-154930			TR 36.881 v0.3.1 for study on latency reduction techniques for LTE as a result of email discussion [91bis#04][LTE/LATRED] and [91bis#04][LTE/LATRED]: changes are based on R2-154929 (revised R2-154744)	0.3.0	0.3.1
2015-10	RAN2#91bis	R2-155008			Agreed TR 36.881 v0.4.0 for study on latency reduction techniques for LTE as a result of email discussion [91#29][LTE/LATRED] and [91bis#04][LTE/LATRED]	0.3.1	0.4.0
2015-11	RAN2#92	R2-156929			TR 36.881 v0.4.1 for study on latency reduction techniques for LTE as a result of RAN2#92 agreements and RAN2 email discussion [92#21][LTE/LATRED] – RAN2 TR 36.881	0.4.0	0.4.1
2015-11	RAN2#92	R2-157181			Agreed TR 36.881 v0.5.0 for study on latency reduction techniques for LTE as a result of RAN2#92 agreements and RAN2 email discussion [92#21][LTE/LATRED] – RAN2 TR 36.881	0.4.1	0.5.0
2016-02	RAN2#93	R2-161869			TR 36.881 v0.5.1 for study on latency reduction techniques for LTE including result of RAN1#84 agreement.	0.5.0	0.5.1
2016-02	RAN2#93	R2-161963			TR 36.881 v0.5.2 for study on latency reduction techniques for LTE as a result of RAN1#84 agreements	0.5.1	0.5.2
2016-02	RAN2#93	R2-161964			Agreed TR 36.881 v0.6.0 for study on latency reduction techniques for LTE as a result of RAN1#84 agreements	0.5.2	0.6.0