Transport Network Design for FrontHaul

Philippe Sehier, Anne Bouillard, Fabien Mathieu, Thomas Deiß

Transport Network Design for FrontHaul

Philippe Sehier*, Anne Bouillard†, Fabien Mathieu† and Thomas Deiß*
*Nokia International †Nokia Bell Labs France
Email: philippe.sehier@nokia.com

***Abstract*—The evolution of LTE and advent of 5G networks increases further the bandwidth requirements for Radio Access Network (RAN). In parallel, the deployment of Centralized RAN architecture raises new challenges on the FrontHaul network. The inflexibility of the legacy Common Public Radio Interface (CPRI) is the primary challenge to Virtualized RAN deployments, and there is currently a strong trend towards the use of packetized transport methods, together with flexible split RAN based architectures. Functional splits within the real-time functions of the RAN have very stringent requirements on latency and jitter. This paper analyzes the jitter produced in the switching nodes of the FrontHaul network, and proposes dimensioning rules.**

I. Introduction

5G aims to enable the deployment of new services with a wide range of requirements, thanks to network virtualization together with more flexible and scalable architectures.

The evolution of LTE and advent of 5G networks increases the bandwidth requirements for the FrontHaul, and the inflexibility of the legacy CPRI is the primary challenge to Virtualized RAN (V-RAN) deployments. The industry is moving towards packet based transport methods to drive the costs down and achieve a better flexibility and scalability.

The most critical challenge is the need to accommodate different transport assets having limited throughputs or high latency. Flexible functional splits aim to enable services with various latency requirements: time critical functions can be located at the network edge, while less time sensitive functions are placed at central locations.

The split options can be classified in two categories: Real-Time (RT) and Non Real-Time. The RT splits raise most of the new challenges, as they are subject to very strict timing requirements, and may become the main dimensioning factor of the Transport Network (TN). The objective of this paper is to characterize the jitter caused by queuing in the switching nodes in the case of RT splits, and deduce how the transport links should be dimensioned to ensure that all but a small bounded fraction of packets are delivered within a target delay.

The rest of the paper is organized as follows: Section II gives a short overview of the FrontHaul architecture. A model of it is presented in Section III and is the basis for our analytical dimensioning rules in Section IV and numerical evaluations in Section V.

II. Assumptions and Background

A. Split architectures in 5G

Split architectures have gained a significant interest in the RAN vendor ecosystem, as well as in the 3rd Generation Partnership Project (3GPP) and various fora. The 3GPP [1] has identified a number of possible split points, as shown in Figure 1.

Each split point option has specific characteristics, like data rate, latency requirements, capability to support advanced features, resource pooling potential, as well as the scalability of throughput with cell loads. 3GPP has estimated that only two split points, one Lower Layer Split (LLS) and one Higher Layer Split (HLS), will be sufficient to fulfill all deployments and use cases:

- LLSs target a very low latency transport, typically under 250 μs. They are based on a split within the RT functions of the RAN, either within the physical layer, or between the Physical (PHY) and Medium Access Control (MAC) layers. They correspond to splits 4 to 8 in Figure 1;
- HLSs have less stringent requirements on latency, and are therefore compatible with most existing transport networks. These are typically splits between the RT and non RT functions of the RAN, and correspond to splits 1 to 3 in Figure 1.

This paper focuses on the analysis of the uplink direction in a LLS where fast Fourier transform, cyclic prefix removal, resource de-mapping and possibly pre-filtering functions reside in the Distributed Unit (DU), while the rest of the PHY functions reside in the Centralized Unit (CU). This split option, referred as option 7-2 in 3GPP [1], offers attractive compression and aggregation gain opportunities.

B. FrontHaul throughput characterization

Two factors can reduce the throughput on the FrontHaul interface: deterministic compression and statistical aggregation. The deterministic frequency domain compression is in the range of $\simeq 5$ to 7 with respect to CPRI for the considered LLS.

The statistical multiplexing gain depends on several factors: number of aggregated cells, traffic correlation and distribution, as well as the characteristics of the load transfer between cells. It can be decomposed in long term (minutes) and short term (few ms) behaviors. The rapid load variations are due to the base station schedulers which generally operate independently in each cell. The present paper focuses on the short term behavior, hence we will assume that these variations are fully de-correlated.

C. Reference RAN architecture

Figure 2 depicts a generic Centralized RAN (C-RAN) architecture. Legacy DUs are interfaced to the FrontHaul network via a CPRI gateway in charge of low layer functions and

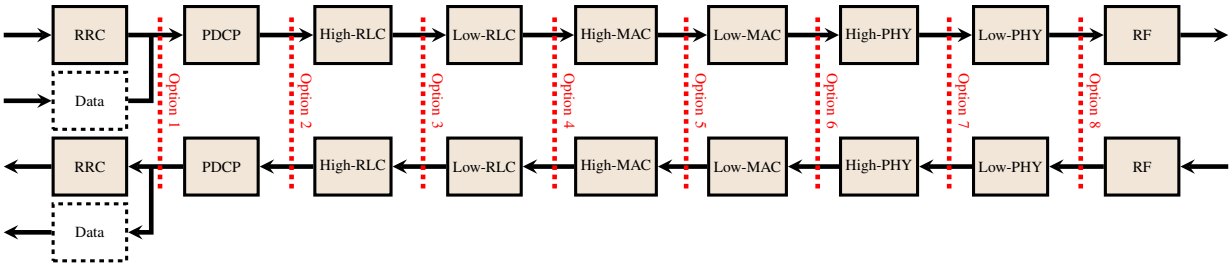


Fig. 1: Possible options for splitting functions between central and distributed units.

packet processing. New generation DUs are natively interfaced to Ethernet and directly connected to the C-RAN. We assume they include the low layers functions.

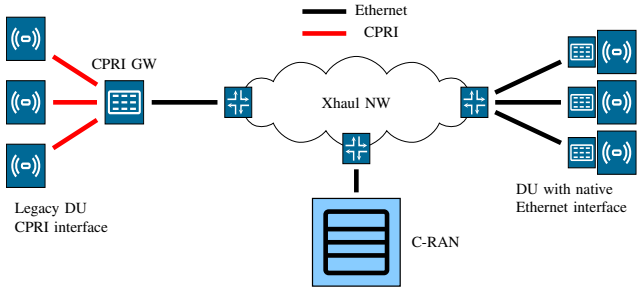


Fig. 2: Overall architecture.

III. FrontHaul modeling

We now detail the main notations and assumptions used for our analysis, which are also summarized in Table I.

A. Simplified architecture

While the FrontHaul network might be composed of several switches, it usually forms a tree directed to the C-RAN. To focus on the impact of multiplexing, we abstract the network as illustrated in Figure 3: N_c cells (DUs) are multiplexed on a single link. The aggregated traffic is processed at a constant rate μ in a single queue.

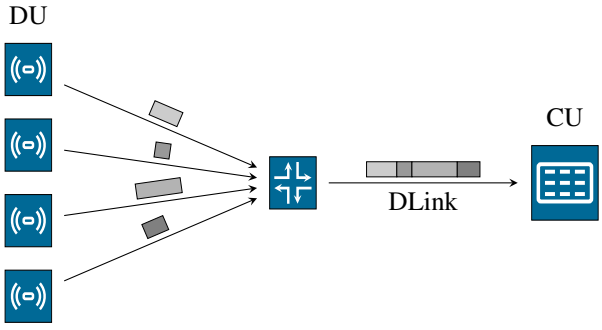


Fig. 3: Simplified reference model.

We assume standard switching mechanisms without any kind of advanced feature like pro-active dropping. The queue length $X_{\max}$ is large enough to be considered infinite (overflow effects are neglected).

B. Packet model

The structure of LTE packets, summarized in Figure 4, is as follows: the elementary transmission block in LTE is a Physical Resource Block (PRB). Each PRB is composed of 168 Resource Elements (REs) over 12 sub-carriers and $m = 14$ symbols. A sub-frame has a duration of $T_s = 1ms$, and can stack up to N_p PRBs in frequency. Following [2], we assume a quantization of 16 bits per RE, so the number of bits per PRB on the FrontHaul interface is $Q = 16 \times 168 = 2688$.

At each sub-frame, each User Equipment (UE) sends a variable number of PRBs to a DU. Each DU gathers the PRBs received (up to N_p) into a packet. Packets are sent in the queue at the symbol time scale: every T_s/m, a *fragment* of one-m^{th} of the packet is added to the queue.

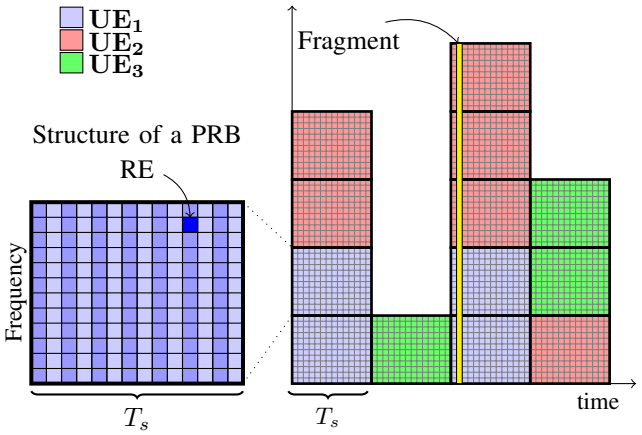


Fig. 4: LTE frame structure and packets formation.

C. Traffic model

We assume an average traffic intensity of λ PRBs per sub-frame in the queue (λ is physically bounded by $\Lambda := N_p N_c$). We suppose for all DUs the same i.i.d. distribution of the number of PRBs sent at each sub-frame. That distribution is *a priori* unknown beyond its mean λ/N_c and its maximum N_p. To determine a performance lower bound, we select a worst-case distribution. It is well-known that for many queuing problems worst-case performance is obtained when the variance is maximized, which gives here the bimodal distribution: at each sub-frame, for each DU, N_p PRBs arrive with probability λ/Λ, and no PRB arrives otherwise.

Symbol	Explanation	Default Value(s)
T_s	Subframe duration	1 ms
m	Number of symbols per T_s	14
N_p	Allocation granularity	10 (PRBs)
N_c	Number of cells	$9, 57$
Λ	Maximal traffic	$N_p N_c$ (PRBs per T_s)
λ	Traffic intensity	0.4Λ, 0.8Λ
$A^{(i)}$	Traffic distribution of cell i	Bimodal: $\mathbf{P}(N_p) = \lambda$ $\mathbf{P}(0) = 1 - \lambda$
μ	System capacity	Integer (PRBs per T_s)
ϵ	Allowed tardy rate	$10^{-3}, 10^{-8}$
δ	Target delay	$0.07T_s \approx T_s/m$
X_n	Queue length at sub-frame n	Integer (PRBs)
$X_{\max}$	Maximum queue length	1000 (PRBs)

TABLE I: Notation and default values.

IV. Analytical approach

The objective is to determine the queue distribution and derive the percentage of packets delivered after some deadline depending on the parameters of our simplified network architecture. Without loss of generality, we assume T_s as the time unit and one PRB as the data unit.

Let $A_n^{(i)}$ be the size of the packet that arrives from cell i at time $n-1$ (that is, for the n-th sub-frame). The total amount of data that arrives at time $n-1$ is $A_n = \sum_i A_n^{(i)}$. As $(A_n^{(i)})_{i,n}$ is assumed i.i.d. on the finite set $\{0, \ldots, N_p\}$, $(A_n)_{n\in\mathbb{N}}$ is also i.i.d. on $\{0, \ldots, \Lambda\}$. We denote $a_k = \mathbf{P}(A_1 = k)$ and remind that $\lambda = \mathbf{E}[A_1]$ is the traffic intensity. At each symbol time of sub-frame n, data of length A_n/m is sent to the link. For the ease of analysis, we assume that the service rate is a whole number: $\mu \in \mathbb{N}$.

Despite the strong local correlation (all the m symbol times of a given sub-frame have the same data size), the system can be seen as a $D/G/1$ queue if observed at the beginning of the sub-frames (just before the arrival of the packets). Let X_n be the size of the queue at time n, just before the arrival of the $n+1$-th packet.

Lemma 1: The stochastic process $(X_n)_{n\in\mathbb{N}}$ satisfies the recursive equation

$$\begin{cases} X_0 = 0 \\ X_{n+1} = \max\left(X_n + A_{n+1} - \mu, 0\right). \end{cases} \tag{1}$$

Under the stability condition $\lambda < \mu$, Eq. (1) defines an ergodic Markov chain $\{X_n\}_{n\in\mathbb{N}}$, that admits a unique stationary distribution that we denote π. The transition matrix $P = (p_{i,j})_{i,j\in\mathbb{N}}$ of this Markov chain is defined by:

$$p_{i,j} = \begin{cases} \sum_{k=0}^{\mu-i} a_k & \text{if } j = 0 \text{ and } i \le \mu, \\ a_{j+\mu-i} & \text{if } j > 0, \\ 0 & \text{otherwise.} \end{cases}$$

The stationary distribution satisfies $\pi = \pi P$ and $\sum_{k\in\mathbb{N}} \pi_k = 1$. An analysis following the lines of [3] shows that the queue length distribution has an exponential decay.

A. Delay experienced by a packet

We now compute the probability that a PRB will not be transmitted on time for a given sub-frame. We denote by δ the maximum delivery time that can suffer fragments of a packet without being considered tardy. We assume for the queuing analysis that $\delta > 1/m$: the target delay is greater than a symbol time. As the last packet introduced in the queue will suffer the largest delay, we consider D_n the delay suffered by the n-th packet of the last cell that was sent at time $n-1$.

Suppose that when a fragment arrives, the queue length is B. Then, the delay experienced by that fragment is B/μ plus the time to send the fragment. A fragment will be tardy when its arrival makes the queue length greater that $\delta\mu$.

When a fragment is tardy, we consider the whole sub-frame of that fragment tardy. With this assumption, a sub-frame is tardy if and only if the first or last fragment of the last cell is tardy. There are different possibilities for the n-th sub-frame to be tardy: A_n must be positive and

- if $X_n > \mu(\delta - \frac{1}{m})$, the last fragment of sub-frame n, which was inserted $1/m$ ago, has a backlog greater than $\delta\mu$: sub-frame n is tardy;
- if $X_n \le \mu(\delta - \frac{1}{m})$:
  - if $A_n > \mu$, then the backlog increases at each fragment of sub-frame n. As the last fragment has a backlog less or equal than $\delta\mu$, the sub-frame is not tardy;
  - if $A_n \le \mu$, then the backlog decreases at each new fragment of the sub-frame, so the n-th sub-frame is tardy if the first fragment is, that is if $X_{n-1} + A_n/m > \delta\mu$.

With $B = \mu(\delta - \frac{1}{m})$, it can be deduced

$$\begin{aligned} \{X_n \ge B\} \cap \{A_n > 0\} &\subseteq \{D_n \ge \delta\} \\ &\subseteq \{X_n \ge B\} \cup \{X_{n-1} + \frac{A_n}{m} \ge \delta\mu \cap A_n \le \mu\} \\ &\subseteq \{X_n \ge B\} \cup \{X_{n-1} + \frac{\mu}{m} \ge \delta\mu\}, \text{ so} \end{aligned}$$

$$\mathbf{P}(X_n \ge B, A_n > 0) \le \mathbf{P}(D_n \ge \delta) \le \mathbf{P}(X_{n-1} \vee X_n \ge B).$$

Applying Eq. (1) and the independence of X_{n-1} and A_n leads to an alternative formulation for the left-hand side:

$$\begin{aligned} &\mathbf{P}(X_n \ge B) - \mathbf{P}(X_n \ge B, A_n = 0) = \\ &\qquad \mathbf{P}(X_n \ge B) - \mathbf{P}(X_{n-1} \ge B + \mu)\mathbf{P}(A_n = 0). \end{aligned}$$

As the queue length distribution has an exponential decay, the last term is negligible. Moreover, $\mathbf{P}(A_n = 0)$ also has an exponential decay with the cell number: the last equation can be approximated by $\mathbf{P}(X_n \ge B)$.

B. Burst of tardy packets

In many systems, bursts of errors have a more detrimental effect than random errors. The effect of bursts can be mitigated by using interleaving techniques, but this cannot be used for the FrontHaul because of the additional delay incurred. It is therefore relevant to analyze the characteristics of tardy packets bursts.

This section analyzes the probability of having a series of tardy packets. The queuing effect induces that a tardy packet at sub-frame n leads to a higher chance to have a tardy packet at

sub-frame $n+1$. In fact, this effect is limited, as the probability to have a series of length r exponentially decays with r.

We focus on the probability of a tardy sub-frame due to a tardy last fragment. As we have just seen, it can be approximated by the probability that the queue length at the end of a sub-frame exceeds B.

For $r \geq 1$, we focus on the quantity

$$\begin{aligned} T_n(r) :&= \mathbf{P}(X_{n+r} \geq B \mid X_n \geq B) \\ &= \sum_{i,j \geq B} \mathbf{P}(X_{n+r} = j \mid X_n = i)\mathbf{P}(X_n = i \mid X_n \geq B). \end{aligned}$$

In the stationary regime, we obtain

$$T(r) = \frac{\sum_{i,j \geq B} (P^r)_{i,j} \pi_i}{\sum_{k \geq B} \pi_k}. \tag{2}$$

To compare with, in the stationary regime, using the same approximation $X_n \geq B$, the probability of a tardy sub-frame is $T(\infty) := \sum_{k \geq B} \pi_k$. We propose to use

$$\frac{T(r)}{T(\infty)} = \frac{\sum_{i,j \geq B} (P^r)_{i,j} \pi_i}{(\sum_{k \geq B} \pi_k)^2} \tag{3}$$

as an estimate of the burstiness impact: it indicates how the odds of a tardy sub-frame are boosted consecutively to the occurrence of an earlier tardy sub-frame.

V. Numerical Evaluation

To characterize the fragment delivery time, we used the following methodology: First we compute the distribution A_n by self-convolving the bimodal (e.g. worst-case) distribution N_c times. Then, we compute the stationary distribution π of the queue by iterating (1) until convergence to the fixed point. Combining A_n and π, and distinguishing the cases $A_n > \mu$, $A_n \leq \mu$ and $A_n = 0$, we compute the delay distribution for sub-frames where at least one PRB was produced. Given some error rate ϵ, we can then deduce the delay δ achieved by a fraction $1 - \epsilon$ of non-empty sub-frames. Note that this evaluation excludes the additional delays caused by the propagation over the transport medium, as well as the processing time in switch nodes. The choice of parameters, inspired by the LTE air parameters, is summarized in Table I.

A. Impact of overprovisoining on delay

Figures 5a and 5b show δ as a function of the relative link capacity μ/Λ, for 2 possible values of λ. We observe the following trends:

- As expected, all other things being equal, targeting a very low error rate ($\epsilon = 10^{-8}$) often translates to a much higher delay;
- The statistical multiplexing of using $N_c = 57$ cells instead of $N_c = 9$ is quite significant: the delay for $N_c = 57, \epsilon = 10^{-8}$ is lower than for $N_c = 9, \epsilon = 10^{-3}$;
- The improvement of overprovisioning after reaching the delay $0.07 \approx 1/m$ is very low. Intuitively, the explanation is that in order to reach $1/m$, we mostly need to contain the effects of the distribution variance on the queue. On the other hand, below $1/m$, the queuing effects become negligible and the delay can be approximated by the time to dispatch a peak fragment, $\Lambda/(m\mu)$, which decreases very slowly with μ.

B. Achieving a target delay

For given target delay δ and error rate ϵ, we can compute the minimal value of μ that achieves δ up to an error rate ϵ. Figure 6 displays the result for several parameters. For cross-comparison, we actually display the relative load λ/μ (the closer to 1, the better).

Figure 6a shows the load as a function of the number of cells for different values of λ and ϵ. As expected, it can be observed that the statistical multiplexing gains increases with the number of cells. This effect is more visible for the low load value ($\lambda = 0.4\Lambda$).

Figure 6b details the impact of the relative traffic intensity λ/Λ. Besides the effects of multiplexing (N_c) and precision (ϵ) already observed, one observes a lower bound that corresponds to $\mu = \Lambda$. It reminds that if the capacity is set for the peak throughput, the delay $1/m$ can always be achieved.

Lastly, Figure 6c evaluates the impact of δ. Again, it appears clearly that targeting a low latency (below $1/m$) has a high cost in term of capacity over-provisioning.

C. Bursts of tardy sub-frames

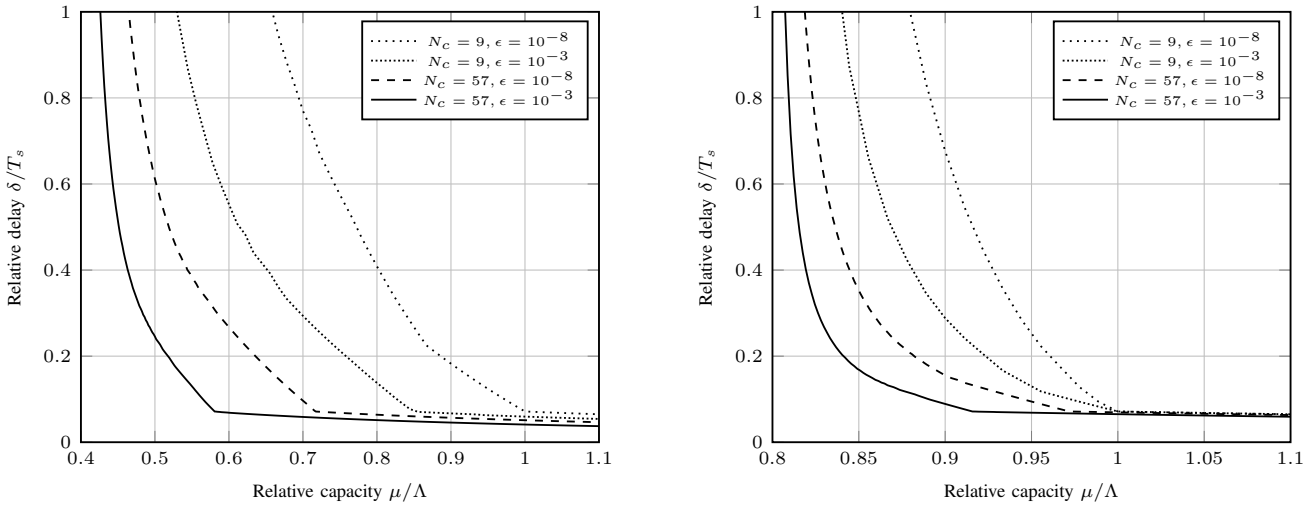
To study the correlation between tardy sub-frames: we fix $\delta = 0.14T_s$[1] and $\epsilon = 10^{-3}$; then, we compute the minimal value of μ for these constraints; finally, we use Equation (3). The results, shown in Figure 7, indicate that the odds of having two consecutive tardy sub-frame are significantly higher than in the hypothesis where tardy sub-frames are fully uncorrelated. On the other hand, the overrisk rapidly decreases and becomes negligible after a few sub-frames. Note that the main impacting parameter is the relative traffic intensity λ/Λ: the closer to 1, the higher the odds of bursts.

VI. Conclusion

Our study confirms that capacity savings on FrontHaul are possible on Uplink with interface option 7.2, while controlling the maximum jitter amplitude. We have analyzed the trade-offs between TN dimensioning and jitter, and derived a methodology to determine the TN link sizes for a predefined latency. These results are based on a number of assumptions relative to the traffic structure which may not be all met in practice. Operators may not take the risk of dimensioning their TN solely based on these results. However, there exist cases where operators could benefit from this analysis. Examples are unexpected traffic ramp up or leased transport resource.

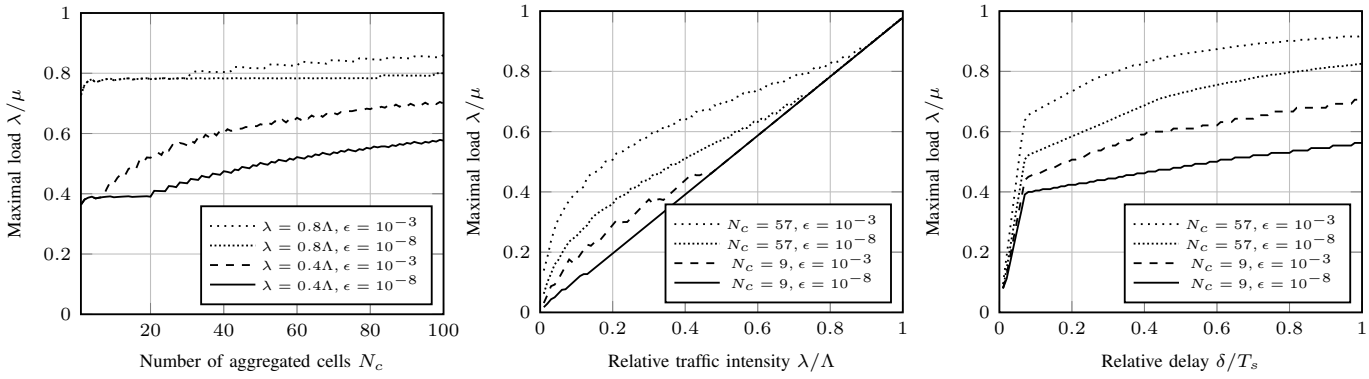
The 5G New Radio (NR) introduces richer numeric settings, and possibly new waveforms in later phases. More stringent latency requirements are also introduced and will create additional constraints on processing times and transport latency for split RAN architectures. The current analysis

[1]The value was increased compared to previous simulations to ensure that $B > 0$ so the assumptions from Section IV are met.



(a) $\lambda = 0.4\Lambda$ (b) $\lambda = 0.8\Lambda$

Fig. 5: Delay achieved by a fraction $1 - \epsilon$ of non-empty subframes with N_c cells aggregated.



(a) Impact of multiplexing ($\delta = 0.07T_s$) (b) Impact of traffic intensity ($\delta = 0.07T_s$) (c) Impact of target delay ($\lambda = 0.4\Lambda$)

Fig. 6: Study of the maximal load that achieves a target delay as a function of the parameters.

could be extended to include the impacts of the 5G NR new possibilities. HLS transport also offers interesting investigation fields with differentiated priority management. Furthermore, the analysis of HLS, LLS, and possibly BackHaul as well as other kinds of background traffics are additional challenges.

Acknowledgement: One author of this paper has been sponsored in part by the project H2020-ICT-2014-2 *5G-Crosshaul: The 5G Integrated fronthaul/backhaul* (671598).

REFERENCES

[1] 3GPP Consortium. Study on new radio access technology: Radio access architecture and interfaces. Technical Report TR-38.801, 3GPP, 2016.

[2] U. Doetsch, M. Doll, H.-P. Mayer, F. Schaich, J. Segel, and P. Sehier. Quantitative analysis of split base station processing and determination of advantageous architectures for LTE. *Bell Labs Tech. J.*, May 2013.

[3] J. F. C. Kingman. A martingale inequality in the theory of queues. *Proceedings of the Cambridge Philosophical Society*, 60:359, 1964.

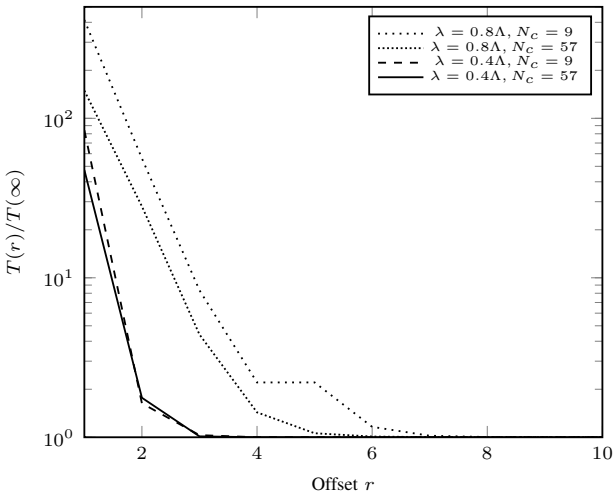


Fig. 7: Increase of the odds of a tardy sub-frames at time r after a tardy sub-frame ($\delta = 0.14T_s, \epsilon = 10^{-3}$).