Reliability and Low Latency: Impact of The Architecture.

Tania Alhajj, Xavier Lagrange

Reliability and Low Latency: Impact of The Architecture.

Tania Alhajj
IMT Atlantique
IRISA, UMR CNRS 6074
F-35700 Rennes, France
tania.alhajj@imt-atlantique.fr

Xavier Lagrange
IMT Atlantique
IRISA, UMR CNRS 6074
F-35700 Rennes, France
xavier.lagrange@imt-atlantique.fr

***Abstract*—Many use cases are meant to be supported by the fifth generation (5G) wireless technology. The one which is occupying the research area for its challenging requirements is the Ultra Reliable Low Latency Communications (URLLC). Hybrid Automatic Repeat reQuest (HARQ) protocol is used to ensure reliability but it induces delay. Furthermore, the transmission in the Radio Access Network (RAN) should be taken into account in the delay budget. In this paper, we jointly analyze the reliability and the delay with two RAN architectures: the legacy one where only one radio unit receives the packet from a terminal and a Centralized-RAN (C-RAN) architecture where several radio units can decode a packet. We propose to combine these approaches in a flexible architecture. The observed enhancement is a division by 850 of the packet erasure rate compared to the legacy architecture with a latency of 3 milliseconds.**

***Index Terms*—5G, URLLC, RAN architecture, C-RAN, flexible architecture.**

I. Introduction

The fifth generation (5G) communication technology is emerging after four previous generations to offer higher rates, lower latency, higher reliability and to serve a larger number of users simultaneously. One of the 5G usage scenarios is the Ultra reliable low latency communications (URLLC). Its importance lies in its use for critical applications. Some of these applications are telemedicine and driverless cars [1]. These applications require very high reliability because they concern human lives. The required reliability ranges from 10^{-3} to 10^{-7} packet error rate (PER) depending on the usage scenario [2]. Since they are critical, a very low latency (between 1 and 10 ms [2]) is required as well. Responding to these two strict requirements while maintaining acceptable network capacity, reasonable energy consumption, and satisfying overall network operation is challenging [3].

Hybrid automatic repeat request (HARQ) is widely known and deployed in wireless networks. It is used to reduce errors and enhance the transmissions reliability [4]. However, packet repetition induces latency. Ensuring high reliability with fewer transmissions requires to reduce the interference. Consequently, this will reduce the network's capacity. Previous studies were made to reduce the latency caused by the HARQ re-transmissions. Cheng and Chen proposed to allocate resources for re-transmissions in advance [5]. Other authors suggested predicting the decoding feedback (acknowledgement (ACK)/negative acknowledgement (NACK)) [6]. This prediction enables earlier re-transmissions. As a result, the delay is shortened.

In the literature, studies were made to respond to URLLC requirements. Some of them were based on varying the transmission time interval (TTI) size and the frame size to reduce the latency [7]. That way, more re-transmissions are allowed before exceeding the accepted delay. Other authors opted the grant free access to cut down the delay generated by the grant based access [8]. Diversity techniques were also studied to achieve high reliability. Time diversity is discarded because it produces latency. In [9] and [10], the multiple reception concept was studied. It is shown that it increases reliability by providing spatial diversity. The authors were interested in studying resource utilization and the network's capacity using this technique.

All these papers consider only the radio interface. The latency perceived by users is also due to the delay in the radio access network. The radio access network (RAN) architecture should thus be included in the study of the latency. Centralized-RAN (C-RAN) is a new network architecture that emerged to be the evolution of the traditional RAN. This centralized architecture is a key element of the 5G technology. It consists of dividing the base station (BS) into two parts: a centralized unit (CU) and a radio unit (RU) [11] (Fig. 1). The connection between these two parts is usually a fiber optic link. High bandwidth, low losses, and interference insensitivity are the interesting advantages of this link. The C-RAN architecture presents many benefits. It helps the operators to reduce the sites' deployment costs. It ensures also the collaboration between different RUs connected to the same pool of CUs. It also provides the ability to choose which BS functions to centralize and which functions to keep implemented locally. This is called the functional split [12] (Fig. 1). Some researchers were interested in the flexibility of

(Author's Accepted Version) (Final version of record is available at: https://doi.org/10.1109/ISCC50000.2020.9219636)

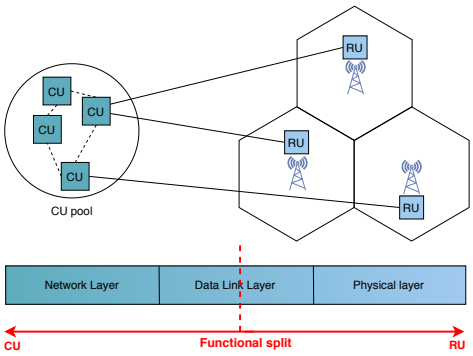


Fig. 1: C-RAN architecture.

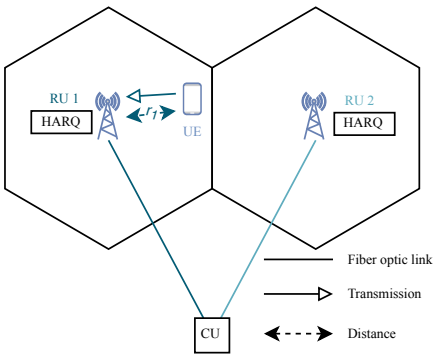


Fig. 2: Architecture A-One receiving BS.

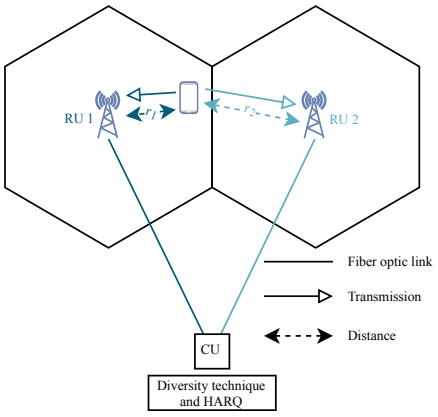


Fig. 3: Architecture B-Multiple receiving BSs ($M = 2$).

the functional split. This flexibility allows switching from an option to another upon the need: whether to reduce energy consumption, to diminish resource utilization, to respond to a latency requirement, or any other need. The adaptive split in [13] is an example.

In this paper, we investigate the impact of the architecture on the latency and on the reliability. We compare the single receiving point to the multiple receiving case while adopting different functional splits. For each architecture, we derive the distributions of the number of transmissions and of the delay. Then we suggest a flexible switch between the two chosen functional split options. Our results are demonstrated analytically and by simulations.

The paper is structured as follows: Section II exposes the two different architectures studied, in addition to the propagation, transmission, and error models. Afterward, the delays for both architectures are detailed in section III. In section IV, we develop the transmissions distributions analytically. Then, in section V, we propose a flexible functional split between the two architectures. Finally, before concluding, we expose and discuss our results in section VI.

II. Architecture and Model

A. Architectures overview

We consider two RAN architectures with different functional splits: A and B.

In A, the medium access control (MAC) layer, responsible for the error correction (using HARQ), is implemented in the RU. This corresponds to functional split option 1 [12]. In this architecture, the data is close to the user [14]. We consider here a user equipment (UE) transmitting and only one receiving RU (Fig. 2). During each transmission, the data is received at the RU. A decoding process occurs. In the case of a failed decoding, the MAC layer in the RU asks for a re-transmission. When the decoding succeeds, the data is transmitted to the CU.

In B, the error correction is triggered in the CU. The MAC layer is therefore implemented in the CU. This is referred to as split option 8 in [12]. Here, we consider that M RUs receive the data from the same UE (Fig. 3). We have a single transmission and multiple receptions. The decoding is based on the M receptions. The coordination between the M RUs is managed by the CU: different receiving RUs send the received signals to the CU where the signals can be combined. In this work, we consider that a reception is successful if the data from at least one of the M RUs is correctly decoded. In case of an error on all the receiving RUs, a re-transmission is requested from the CU.

In A and B, we consider that a CU is connected via fiber optic links to multiple RUs located at the tower site. We consider these RUs to be equidistant from the CU. We define θ, the propagation delay between different RUs and the CU that they're connected to:

$$\theta = \frac{\rho}{v} \tag{1}$$

where ρ is the distance between the CU and the RUs connected to it and v is the velocity of light inside a fiber ($v = 2 \times 10^8$ m/s). Then, for the same CU, θ is the same for all RUs.

B. System model

The channel model is considered as Cost-231 Hata [15]. The received power depends on the path-loss and the fading:

$$P_\mathrm{r} = P_\mathrm{t} \left(\frac{r_0}{r}\right)^\alpha \chi \tag{2}$$

where P_r is the received power, P_t the transmitted power, r_0 is a constant, r the distance between the transmitter and the receiver (the UE and the BS respectively in our case), α the

path-loss exponent and χ is an exponential random variable representing the fading with mean = 1. The shadowing is not taken into consideration.

The transmissions in both architectures are in the uplink (UL) direction. We consider successive transmissions of data packets. All transmitted packets have the same size. We consider that the transmission duration of a packet or an ACK/NACK is 1 TTI. Thereby, we have $T_A = T_D$ with T_A being the transmission duration of the ACK/NACK and T_D the transmission duration of a packet of data. We consider a perfect downlink (DL), i.e. ACKs and NACKs are always successfully received.

We consider the background noise N as a constant. Parameter N includes the noise and the interference. The error in decoding is related to the signal to noise ratio (SNR) which is denoted by γ:

$$\gamma = \frac{P_{\mathrm{r}}}{N}. \tag{3}$$

With the previous considerations, we define $\overline{\gamma}$, the average SNR as:

$$\overline{\gamma} = \frac{P_{\mathrm{t}}}{N}\left(\frac{r_0}{r}\right)^{\alpha}. \tag{4}$$

The PER is approximated and calculated as a function of the SNR similarly to [16]:

$$h(\gamma) = \begin{cases} 1 & \text{if } 0 < \gamma < \gamma_M \\ ae^{-g\gamma} & \text{if } \gamma \geq \gamma_M \end{cases} \tag{5}$$

where a and g are parameters that depend on the modulation and coding scheme (MCS) mode and $\gamma_M = \frac{\ln a}{g}$.

III. Delays

The reliability is achieved by the HARQ chase combining (CC) mechanism in architecture A. For architecture B, both HARQ and spatial diversity are used to achieve reliability. Now, to compute the delay for each architecture, we expose its components. We assume that the processing duration is negligible. We consider that the propagation delay is absorbed by the guard time of the slot. We also suppose that we have a very high rate on the fiber link. Then, the transmission duration over the fiber is negligible. In the following, the number of transmissions is denoted by l: $l-1$ failed transmissions and then a successful one.

For architecture A, we define the cycles duration in both cases: good and bad decoding. In Fig. 4, d_{1f} denotes the delay of 1 cycle in which the decoding fails:

$$d_{1f} = T_D + T_A. \tag{6}$$

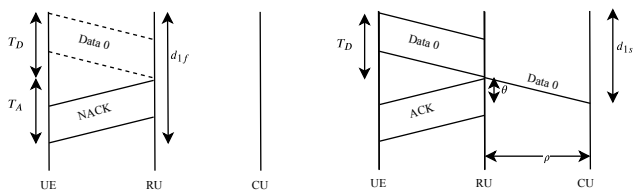
Parameter d_{1s} represents the delay of 1 cycle where the decoding succeeds:

$$d_{1s} = T_D + \theta. \tag{7}$$

Therefore, the total delay produced by l transmissions is:

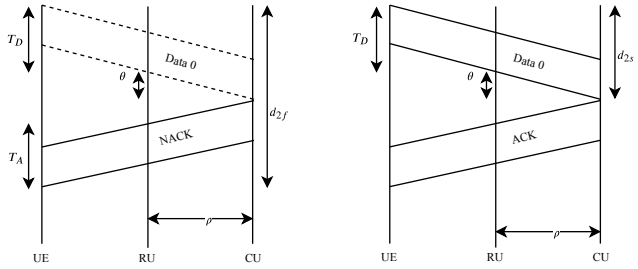
$$d_1 = (l-1)(T_D + T_A) + T_D + \theta. \tag{8}$$

For architecture B, the cycles delays are shown in Fig. 5. The delays illustrated in this figure are considered between



(a) (b)

Fig. 4: 1 cycle delay (a) failure and (b) success case (A).



(a) (b)

Fig. 5: 1 cycle delay (a) failure and (b) success case (B).

a UE and one BS ($\mathrm{BS_i}$). Let d_{2f} and d_{2s} denote the delay of 1 cycle with an erroneous and a successful decoding respectively. Then, we have:

$$d_{2f} = T_D + 2\theta + T_A \tag{9}$$

and

$$d_{2s} = T_D + \theta. \tag{10}$$

The total delay generated by l transmissions in B is:

$$d_2 = (l-1)(T_D + 2\theta + T_A) + T_D + \theta. \tag{11}$$

IV. Analytic formulation

The delay detailed in III is a function of l, the number of transmissions. Consequently, the distribution of l is needed to know the delay and its distribution.

The UEs are considered to be uniformly distributed in an hexagonal cell of radius R_c. The RUs are at the centers of the cells. In [17], the probability of having more than k transmissions as a function of $\overline{\gamma}$ was computed when the PER is given by (5):

$$\begin{aligned} \mathbb{P}(l > k/\overline{\gamma}) &= \Gamma_l(k, X) \\ &+ e^{-X} \sum_{i=0}^{k-1} \frac{(X)^i}{i!} \frac{\Gamma(\frac{1}{g\overline{\gamma}})}{(g\overline{\gamma})^{k-i+1}\Gamma(\frac{1}{g\overline{\gamma}} + k - i + 1)} \end{aligned} \tag{12}$$

where $\mathbb{P}(l > k/\overline{\gamma})$ is the probability of having more than k transmissions for a given $\overline{\gamma}$ in a fading channel, $X = \frac{\gamma_M}{\overline{\gamma}}$, and $\Gamma_l(b, x) = \frac{1}{\Gamma(b)} \int_0^x t^{b-1} e^{-t} \mathrm{d}t$ is the lower incomplete normalized gamma function. The probability distribution function (PDF) of the number of transmissions (l) is obtained by:

$$\mathbb{P}(l = k/\overline{\gamma}) = \mathbb{P}(l > k - 1/\overline{\gamma}) - \mathbb{P}(l > k/\overline{\gamma}). \tag{13}$$

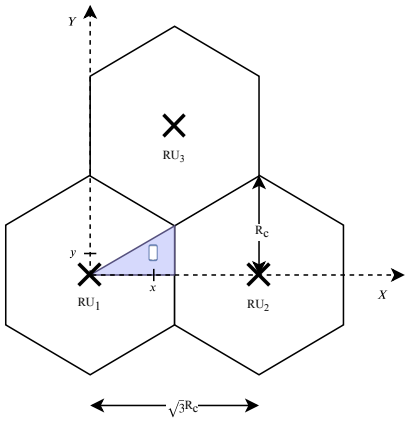


Fig. 6: Zone of study in the hexagonal cell.

For a fixed P_t and N, $\overline{\gamma}$ varies with the distance between the transmitter and the receiver.

For symmetry reasons, our study is limited to the shadowed area of Fig. 6. We consider a UE at the position (x, y) relative to RU_1. We have $0 < x < \frac{\sqrt{3}}{2}R_c$ and $0 < y < \frac{x}{\sqrt{3}}$.

A. Architecture A

In A, the UE in position (x, y) has a corresponding $\overline{\gamma}(x, y) = \frac{P_t}{N}\left(\frac{r_0}{\sqrt{x^2+y^2}}\right)^{\alpha}$. Thus, for each UE, we have a distribution of l, the number of transmissions from (12) and (13). In order to get a distribution of l, the distribution depending on the position should be averaged over all the studied surface:

$$\mathbb{P}(l > k) = \frac{8}{\sqrt{3}R_c^2}\int_0^{\frac{\sqrt{3}}{2}R_c}\int_0^{\frac{x}{\sqrt{3}}} A(x, y)\mathrm{d}y\mathrm{d}x \tag{14}$$

where $A(x, y) = \mathbb{P}\left(l > k/\overline{\gamma}(x, y)\right)$.

B. Architecture B

In B, for $M = 2$, the UE in position (x, y) relative to RU_1, is in position $(\sqrt{3}R_c - x, y)$ relative to RU_2. The average SNR relative to the second BS is $\overline{\gamma}(\sqrt{3}R_c - x, y)$. A transmission is considered erroneous if its reception fails at the 2 RUs. So, the probability of having more than k transmissions is:

$$\mathbb{P}(l > k) = \frac{8}{\sqrt{3}R_c^2}\int_0^{\frac{\sqrt{3}}{2}R_c}\int_0^{\frac{x}{\sqrt{3}}} A(x, y)A(\sqrt{3}R_c - x, y)\mathrm{d}y\mathrm{d}x. \tag{15}$$

For $M = 3$, the UE is in position $(\frac{\sqrt{3}R_c}{2} - x, \frac{3R_c}{2} - y)$ relative to the third RU. Similarly to (15), we get:

$$\mathbb{P}(l > k) = \frac{8}{\sqrt{3}R_c^2}\int_0^{\frac{\sqrt{3}}{2}R_c}\int_0^{\frac{x}{\sqrt{3}}} A(x, y) A(\sqrt{3}R_c - x, y)A\left(\frac{\sqrt{3}R_c}{2} - x, \frac{3R_c}{2} - y\right)\mathrm{d}y\mathrm{d}x. \tag{16}$$

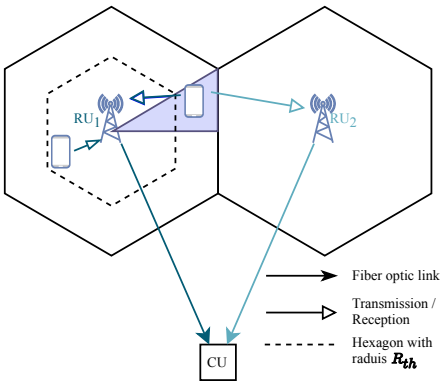


Fig. 7: Zone delimitation to switch between A and B.

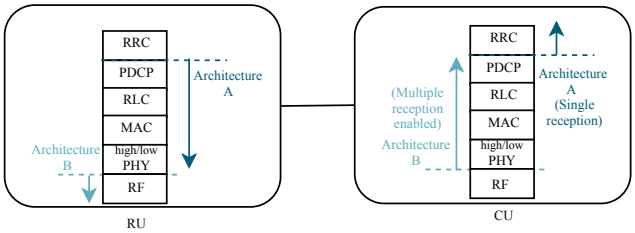


Fig. 8: Flexible split between A and B.

V. Flexible switch between A and B

After presenting both architectures, we propose to dynamically switch between them. In fact, when the UE is too close to the first RU, there is no need to let the second RU receive from it. In this case, architecture A and single reception are considered. On the other side, when the UE is near the cell edge, we switch to architecture B and we enable multiple receptions. We define R_{th}, the radius of the limit zone outside which we allow two RUs to receive using architecture B (Fig. 7). If the UE is inside the mentioned zone, architecture A and single reception are enabled. This switch is done by the flexible functional split. The CU measures x and chooses the adopted architecture accordingly. For each transmission to/from the considered UE, each unit encodes/decodes the data until the corresponding layer (dashed lines in Fig. 8) and transmits to the other unit. The switching algorithm is the following:

Algorithm 1: Flexible switch between A and B

```
if 0 < x < (√3/2) R_th then
    consider architecture A;
else if x < (√3/2) R_c then
    consider architecture B;
    enable multiple receptions;
end
```

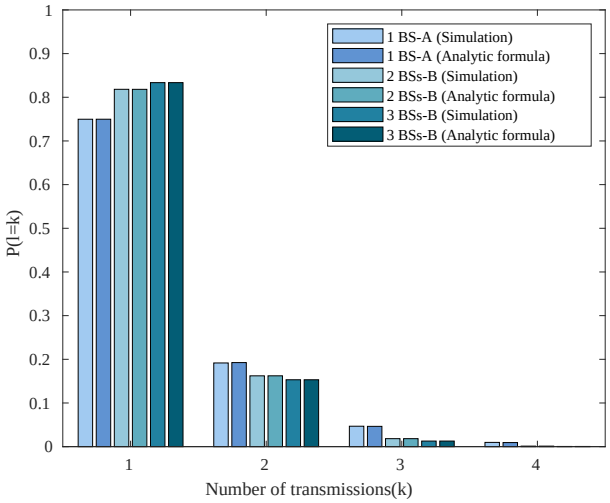


Fig. 9: Probability distribution of the number of transmissions for A and B.

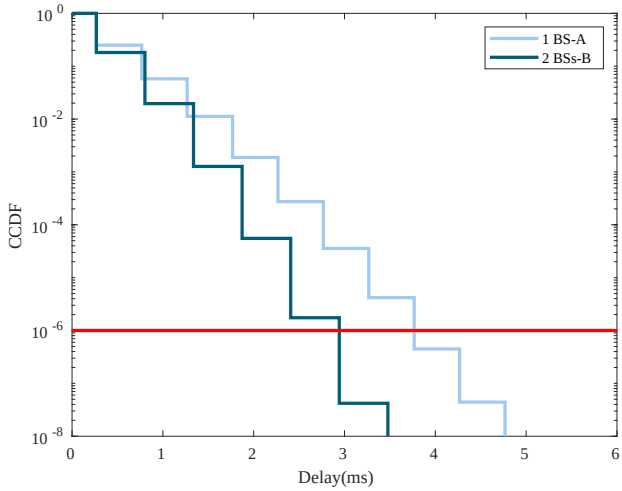


Fig. 10: Delay CCDF for A (with 1 BS) and B (with 2 BSs).

For the flexible split with $M = 2$, we have:

$$\mathbb{P}(l > k) = \frac{8}{\sqrt{3}R_c^2}\left(\int_0^{\frac{\sqrt{3}}{2}R_{th}} \int_0^{\frac{x}{\sqrt{3}}} A(x,y)\mathrm{d}y\mathrm{d}x + \int_{\frac{\sqrt{3}}{2}R_{th}}^{\frac{\sqrt{3}}{2}R_c} \int_0^{\frac{x}{\sqrt{3}}} A(x,y)A(\sqrt{3}R_c - x, y)\mathrm{d}y\mathrm{d}x \right). \quad (17)$$

VI. Results

In this section we present our analytic and simulations results. Our simulations follow the analytic steps. The propagation and error models from (2) and (5) are used. The parameters values for our numerical calculations and our simulations are summarized in table I.

TABLE I: Parameters values.

Symbol	Parameter	Values
P_t (dBm)	UE's transmission power	23
r_0 (m)	Reference distance	0.2
α	Path-loss exponent	3.38
N (dBm)	Noise power	-116.45
a [16]	Parameter depending on the MCS	274.7
g [16]	Parameter depending on the MCS	7.993
R_c (km)	Cell radius	3.2
M	Number of receiving BSs	1 for A, 2/3 for B
ρ (Km)	CU-RUs distance	3.5
T_D (ms)	Data transmission duration	1 TTI=0.25 [a]
T_A (ms)	ACK/NACK transmission time	1 TTI=0.25 [a]

[a] Numerology 2 of the 5G new radio (NR) [18].

A. Comparison of the two architectures

Fig. 9 shows the similarity between the mathematical computation and the simulation results for $\mathbb{P}(l = k)$. It also compares the single reception (with A) to the multiple receptions (with B). It is shown that for the case of 2 receiving BSs, we have higher chances to get a successful decoding from the first transmission. For further transmissions, the probability is lower for the case of 2 BSs. Thereby, in B, the number of transmissions needed to get a successful decoding is reduced. No significant improvement is observed for receiving from a third BS. This is because the third BS is too far. For this reason, we limit our study to $M = 2$.

The delay depends on the number of transmissions (l). We can get the delay's distribution based on the distribution of l. Fig. 10 illustrates the complementary cumulative distribution function (CCDF) of the delay. Having the number of transmissions with a certain probability, we get the latency's distribution. If we define the reliability as receiving the data successfully within a certain duration, then this CCDF represents the outage probability. For example, if we want to assure a delay less than 2 ms, we have higher reliability by a factor of 850 in the case of 2 BSs (architecture B) compared to 1 (architecture A). This improvement is larger if we allow higher delays. If we desire to get an outage probability of 10^{-6}, we can see in Fig.10 that we get lower latency with B (2.9 ms for B compared to 3.8 ms latency for A). We can notice that distancing the CU from the RUs can add more delay. This additional delay affects B more than A (which is seen in (8) and (11)). This leads the delay's CCDF of B to approach the CCDF of A. The difference between the 2 delays (for A and B) at 10^{-6} PER is 0.83 ms. If we let $\rho = 20$ km, (8) and (11) give the same results for the mentioned PER. A distance higher than 20 km increases more the delay in B. That way, we observe higher delays in architecture A compared to architecture B for the same PER.

B. Flexible C-RAN architecture

Fig. 11 illustrates the distribution of l, the number of transmissions, for different R_{th}. We note that $R_{th} = R_c$ is equivalent to the case of adopting architecture A with one receiving BS and $R_{th} = 0$ corresponds to architecture B with two receiving BSs. The threshold is chosen when we start getting the same performance as if we are receiving from 2 BSs. Accordingly, in our case, we choose $R_{th} = 0.6R_c$. This is also shown in Fig.12, where we have the same performances

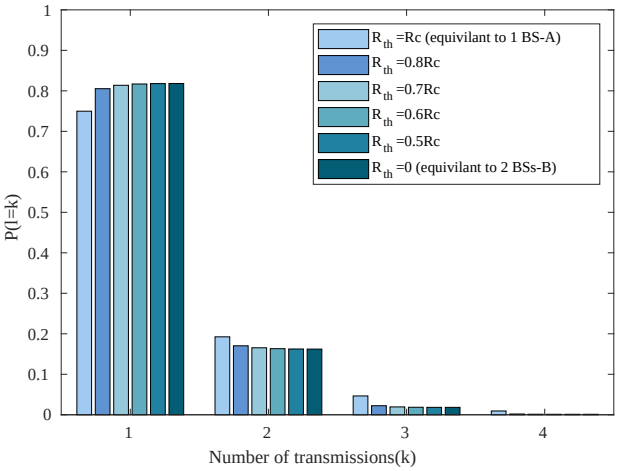


Fig. 11: Probability distribution of the number of transmissions for different R_{th}.

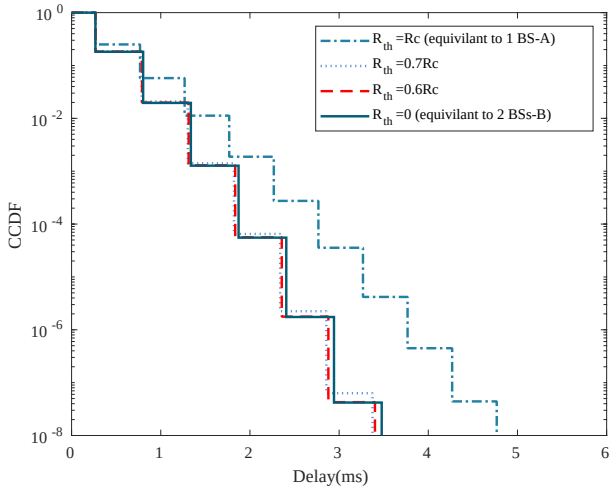


Fig. 12: Delay CCDF with different R_{th}.

for the case of $R_{th} = 0.6R_c$ and $R_{th} = 0$. The delays achieved by the switching algorithm are slightly lower than the case where B is only implemented (2.34 ms and 2.4 ms respectively for 10^{-5} PER). The maximum observed difference is 0.13 ms. By flexibly splitting the BS's functions, we save approximately 40% of the use of the second RU.

VII. Conclusion

In this paper, we took two different C-RAN architectures. In the first one, we chose the re-transmissions to be triggered in the RU and we adopted single reception. In the second one, the re-transmission decision was made in the CU and we adopted multiple receptions. It was shown that lower delays can be achieved, for the same reliability, when receiving from more than one BS. These lower delays are reached even if we have higher round trip time (RTT) per transmission. The spatial diversity reduces the number of transmissions needed to get a successful decoding. As a result, lower delays can be reached. We proposed to dynamically switch between A and B. We proved that by enabling the reception from two BSs just when needed, we can reach the aimed target: lower latency and higher reliability.

In future work, we can consider a combining diversity technique that provides more reliability such as maximum ratio combining (MRC).

References

[1] "New services and applications with 5G ultra-reliable low latency communications," 5G Americas, White Paper, November 2018.

[2] Z. Li, M. A. Uusitalo, H. Shariatmadari, and B. Singh, "5G URLLC : Design challenges and system concepts," in *2018 15th International Symposium on Wireless Communication Systems (ISWCS)*, August 2018, pp. 1–6.

[3] B. Soret, P. Mogensen, K. I. Pedersen, and M. C. Aguayo-Torres, "Fundamental tradeoffs among reliability, latency and throughput in cellular networks," in *2014 IEEE Globecom Workshops (GC Wkshps)*, December 2014, pp. 1391–1396.

[4] C. J. Le Martret, A. Le Duc, S. Marcille, and P. Ciblat, "Analytical performance derivation of hybrid ARQ schemes at IP layer," *IEEE Transactions on Communications*, vol. 60, no. 5, pp. 1305–1314, May 2012.

[5] K. Cheng and J.Chen, "Dynamic pre-allocation HARQ (DP-HARQ) in IEEE 802.16j mobile multihop relay (MMR)," in *2009 IEEE International Conference on Communications*, June 2009, pp. 1–6.

[6] G. Berardinelli, S. R. Khosravirad, K. I. Pedersen, F. Frederiksen, and P. Mogensen, "Enabling early HARQ feedback in 5G networks," in *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, May 2016, pp. 1–5.

[7] C.-P. Li, J. Jiang, W. Chen, T. Ji, and J. E. Smee, "5G ultra-reliable and low-latency systems design," 2017, pp. 1–5.

[8] N. H. Mahmood, R. Abreu, R. Böhnke, M. Schubert, G. Berardinelli, and T. H. Jacobsen, "Uplink grant-free access solutions for URLLC services in 5G new radio," in *2019 16th International Symposium on Wireless Communication Systems (ISWCS)*, August 2019, pp. 607–612.

[9] T. H. Jacobsen, R. Abreu, G. Berardinelli, K. I. Pedersen, I. Z. Kovács, and P. Mogensen, "Multi-Cell reception for uplink grant-free ultra-reliable low-latency communications," *IEEE Access*, vol. 7, pp. 80 208–80 218, 2019.

[10] N. H. Mahmood, A. Karimi, G. Berardinelli, K. I. Pedersen, and D. Laselva, "On the resource utilization of multi-connectivity transmission for URLLC services in 5G new radio," in *2019 IEEE Wireless Communications and Networking Conference Workshop (WCNCW)*, April 2019, pp. 1–6.

[11] C. I, J. Huang, R. Duan, C. Cui, J. Jiang, and L. Li, "Recent progress on C-RAN centralization and cloudification," *IEEE Access*, vol. 2, pp. 1030–1039, August 2014.

[12] L. M. P. Larsen, A. Checko, and H. L. Christiansen, "A survey of the functional splits proposed for 5G mobile crosshaul networks," *IEEE Communications Surveys Tutorials*, vol. 21, no. 1, pp. 146–172, Firstquarter 2019.

[13] Y. Alfadhli, M. Xu, S. Liu, F. Lu, P. Peng, and G. Chang, "Real-time demonstration of adaptive functional split in 5g flexible mobile fronthaul networks," in *2018 Optical Fiber Communications Conference and Exposition (OFC)*, March 2018, pp. 1–3.

[14] 3GPP, "Study on new radio access technology; radio access architecture and interfaces," 3rd Generation Partnership Project (3GPP), TR 38.801, March 2017.

[15] 3GPP, "Evolved universal terrestrial radio access (e-utra); radio frequency (rf) system scenarios," 3rd Generation Partnership Project (3GPP), TR 36.942, July 2018.

[16] Q. Liu, S. Zhou, and G. Giannakis, "Cross-layer combining of adaptive modulation and coding with truncated ARQ over wireless links," *Wireless Communications, IEEE Transactions*, vol. 3, pp. 1746 – 1755, October 2004.

[17] X. Lagrange, "Throughput of HARQ protocols on a block fading channel," *Communications Letters, IEEE*, vol. 14, pp. 257 – 259, April 2010.

[18] 3GPP, "5G NR physical channels and modulation," 3rd Generation Partnership Project (3GPP), TS 38.211, July 2018.