Impact of the RAN Architecture and Macro-diversity Techniques on Latency

Tania Alhajj, Xavier Lagrange

Impact of the RAN Architecture and Macro-diversity Techniques on Latency

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*—In fifth generation (5G) wireless technology, the centralization of some base station functions will be possible using the new radio access network (RAN) architecture: centralized-RAN (C-RAN). The most challenging type of service to be served by 5G is ultra reliable low latency communication (URLLC), which requires high reliability and low latency simultaneously. In this paper, we compare three RAN architectures. The first has all the processing close to the user and involves single reception. The second and the third architectures, have some centralized BS functions, allowing multiple-cell coordination and thus multiple reception points. We study the impact of architecture and macro-diversity both on the reliability of uplink (UL) packet transmission and on latency.**

***Index Terms*—RAN architecture, Macro-diversity, URLLC, 5G.**

I. Introduction

Multiple applications with different requirements have appeared over time. Three types of service are offered by fifth-generation (5G) wireless networks: enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and ultra-reliable low latency communications (URLLC). The first two types of service are an extension of the previous existing wireless network generations. The new type is the URLLC and it is meant to offer highly reliable communication throughout a very short duration. The requested reliability is typically 10^{-7} packet error rate (PER) and has to be reached within a very short one-way latency that is estimated to be 1 ms [1]. Important applications require URLLC, including medical telesurgeries, intelligent transportation, and industrial automation [2]. Researchers are facing challenges by trying to provide reliability and delay demands for novel applications. They are investigating in this direction to answer the delay and reliability demands. However, their studies have also been conducted to maintain the optimization of other network components, such as resource usage, processing capacity, and energy use.

Although additional delays are generated by re-transmissions, their use is unavoidable to guarantee an ultra-reliable communication. That's why researchers are finding use cases of hybrid automatic repeat request (HARQ) where both high reliability and low latency requirements are fulfilled. One example is predicting the decoding result to have earlier feedback [3]. In our study, we consider the impact of macro-diversity on the reliability to optimize the HARQ re-transmission delay.

Centralized-radio access network (C-RAN) is an extension of the existing RAN. It has the possibility of centralizing a number of baseband functions determined by a functional split (FS) [4]. The main components of this architecture are the baseband unit (BBU) and the remote radio head (RRH), known as radio units (RU). The BBU is usually centralized (centralized unit (CU)) and connected via the fronthaul to one or many RRHs. One of the main benefits of C-RAN is cooperation capability. The pool of CUs, implemented in a centralized location, allows communication between different cells served by different RUs. Previous studies have demonstrated the effect of the C-RAN architecture on URLLC [5]. In this study, we show the impact of different FSs on both latency and reliability. Unlike previous studies, we focus on the latency produced over the access network.

Common public radio interface (CPRI) connection protocol was used for many fronthaul connections in long term evolution (LTE). With 5G emerging, higher traffic is expected on the fronthaul link. Also, with the functions distribution between the CU and the RU, the load on the fronthaul appears to be very variable. For some splits, fronthaul traffic scales with the number of users. For others, it scales with the number of multiple input multiple output (MIMO) antennas. Those reasons with other ones were behind the specification of an ethernet-based CPRI (eCPRI) [6]. CPRI uses constant bit rate transmissions while eCPRI is packet-based and varies with the actual payload. Thus, eCPRI avoids resource waste and offers more flexibility. For those reasons, the load in our study is carried on a fiber optic link fronthaul based on eCPRI link protocol.

In a previous study [7], we studied the impact of two RAN architectures on latency and reliability. In this paper, we compare three RAN architectures, illustrated in Fig. 1. In architecture A, only one RU receives the transmitted packet, decodes it, and processes the HARQ process. In architectures B1 and B2, most of the processing is centralized and several RUs receive the data packet. In these two architectures, the

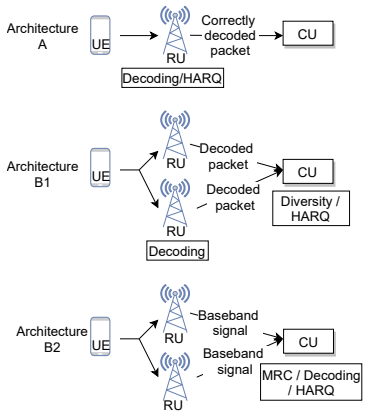

Fig. 1: Architectures A, B1, and B2.

HARQ process is centralized. The main difference is the location of the decoding process. For architecture B1, the decoding is done in the RU, whereas for architecture B2 it is centralized. We compare two combining techniques. In architecture B1, a reception is considered good when at least one of the RUs is able to correctly decode the packet. In architecture B2, the combining technique is the maximum ratio combining (MRC), in which all the received signals are summed in the CU then decoded. In [7], we only assumed a distance-based path loss. In this paper, we consider a more realistic channel model that includes shadowing. The impact of shadowing is important, because the nearest base station (BS) is not always the best receiving BS. The rate on the fronthaul is considered finite in this study, unlike in [7] where it was considered infinite.

The remainder of the paper is structured as follows: in Section II we elaborate the model used. Section III exposes the architectures used. The delays of each architecture are detailed in Section IV. In Section V we expand our analytic formulation of the problem for different cases. We show our simulation and analytic results in Section VI, and we sum up in Section VII.

II. Model

For the sake of completeness of the article, we recall the model used in [7]. We consider a hexagonal cell network where the user equipment (UE) is uniformly distributed. We consider the BS to be split into two units: RU and CU. The RUs are omnidirectional and implemented in the center of each hexagonal cell of radius R_c. The CU, connected to multiple RUs, is implemented at an equal distance ρ from these RUs. RU and CU are connected through a packet-based eCPRI transport-network, reduced to a simple fiber optic link to ensure low latency. The propagation velocity over the fiber link is 2×10^8 m/s. Therefore, the propagation delay θ between the CU and the corresponding RUs is constant.

The propagation model is COST-231 Hata [8]:

$$P_r = P_t \left(\frac{r_0}{r}\right)^{\alpha} e^{\xi_c+\xi_s}\chi, \tag{1}$$

where P_r is the received power, P_t the transmitted power, r_0 a constant reference distance, r the distance between the UE and the BS, α the path-loss exponent, χ an exponential random variable (r.v.) representing fading with mean $= 1$, ξ_c and ξ_s two normal r.v. representing the common and the specific part of the correlated shadowing, respectively: $\xi_c \sim N\left(0, \left(\frac{\sigma_{c_{\text{dB}}}\ln(10)}{10}\right)^2\right)$ and $\xi_s \sim N\left(0, \left(\frac{\sigma_{s_{\text{dB}}}\ln(10)}{10}\right)^2\right)$. The shadowing correlation coefficient is $\delta = \frac{\sigma^2_{c_{\text{dB}}}}{\sigma^2_{\text{dB}}}$, with $\sigma^2_{\text{dB}} = \sigma^2_{c_{\text{dB}}} + \sigma^2_{s\text{dB}}$.

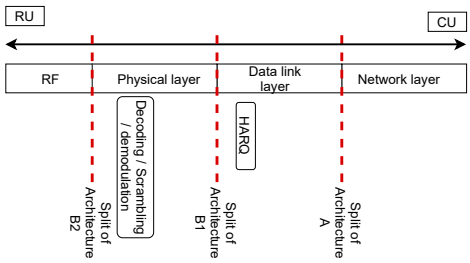

Fig. 2: Functional splits of architectures A, B1, and B2.

We consider two connection types between the UE and the serving RU. In the first one, the UE is connected to the nearest RU. This connection is not the optimal one due to the shadowing effect. In the second one, the UE is connected to the best BS, which is the case of an ideal network. In fact, due to the hysteresis effect, the UE is not always connected to the best BS.

Let us consider the transmissions of data packets in the uplink (UL) direction. The downlink (DL) is considered error free: no losses on the feedback channel. The error model is taken from [9], where the authors computed the PER as a function of the signal-to-noise ratio (SNR):

$$h(\gamma) = \begin{cases} 1 & \text{if } 0 < \gamma < \gamma_M \\ ae^{-g\gamma} & \text{if } \gamma \geq \gamma_M \end{cases} \tag{2}$$

where γ is the SNR, a and g are parameters that depend on the modulation and coding scheme (MCS) mode, and $\gamma_M = \frac{\ln a}{g}$. HARQ with chase combining (HARQ-CC) is used for error correction. The UE transmits a packet. The receiver replies with an acknowledgement (ACK) if the packet is correctly decoded. Otherwise, a negative ACK (NACK) is sent. In the case of an erroneous decoding, another round of HARQ starts again. At the receiver, all the replicas of the packets are saved and combined with the new receptions for decoding. The process is repeated until decoding is successful.

III. Architecture overview

The three architectures laid out in this section are illustrated in figures 1 and 2. They involve a single transmission and M receptions. For architecture A, only one RU receives the signal transmitted by the UE: $M = 1$. For architectures B1 and B2, we have multiple receptions: $M > 1$. We assume that data packets are processed at the CU, which for example includes mobile edge computing functions.

In architecture A, all the processing is done in the RU. This is split A in [6]. The RU receives the packet and decodes it.

If the decoding is successful, the packet is sent to the CU. In case of error, a re-transmission is triggered by the RU.

In architecture B1, each RU receiving the transmitted packet decodes it and forwards it to the CU. This is split D in [6]. In the CU, the redundancy of different decoded packets from different RUs is removed. An error happens if the M RUs fail to correctly decode the packet. In case of error, the medium access control (MAC) layer in the CU asks for re-transmission.

In architecture B2, each RU receiving the transmitted packet forwards it to the CU, where decoding takes place. This is the traditional FS detailed in [10]. In the CU, the signals are combined by the MRC technique. The SNR of the M signals are summed. Then, the decoding process takes place. If an error occurs, a re-transmission is initiated from the CU.

IV. Delay components

In this section, we provide the delay components for architectures A, B1, and B2. The calculated delay includes propagation and transmission duration over the radio interface and over the fronthaul until the CU. The processing delay is not taken into consideration.

A. Architecture A

We define cycles duration in both cases: good and bad decoding. In Fig. 3, $d_{\mathrm{A,f}}$ denotes the delay of one cycle with bad decoding in architecture A:

$$d_{\mathrm{A,f}} = T_{\mathrm{D,R}} + T_{\mathrm{A,R}} + 2\frac{r}{c}, \tag{3}$$

where $T_{\mathrm{D,R}}$ and $T_{\mathrm{A,R}}$ are data and ACK/NACK transmission delay, respectively, over the radio interface, and r/c the propagation delay over the radio interface. For a good reception, the delay of one cycle is:

$$d_{\mathrm{A,s}} = T_{\mathrm{D,R}} + \frac{r}{c} + T_{\mathrm{D,FH}} + \theta, \tag{4}$$

where $T_{\mathrm{D,FH}}$ is the data transmission duration over the fronthaul. In architecture A, a correctly decoded packet in the RU is transmitted over the fronthaul towards the RU. Thus, the transmission over the fronthaul duration is related to the packet size (L_{P}) and its corresponding headers size:

$$T_{\mathrm{D,FH}} = \frac{L_{\mathrm{P}} + L_{\mathrm{H_{ecpri}}} + L_{\mathrm{H_{TN}}}}{C_{\mathrm{FH}}}, \tag{5}$$

where $L_{\mathrm{H_{ecpri}}}$ is the eCPRI header size, $L_{\mathrm{H_{TN}}}$ the transport-network layer headers size, and C_{FH} the maximum throughput over the fronthaul. So, the total delay produced by l transmissions ($l-1$ failed and one successful), for architecture A is:

$$d_{\mathrm{A}} = (l-1)d_{\mathrm{A,f}} + d_{\mathrm{A,s}}. \tag{6}$$

B. Architecture B1

The cycles delays are shown in Fig. 4: $d_{\mathrm{B1,f}}$ denotes the delay of one cycle with bad decoding in architecture B1:

$$d_{\mathrm{B1,f}} = T_{\mathrm{D,R}} + T_{\mathrm{A,R}} + 2\frac{r}{c} + T_{\mathrm{D,FH}} + T_{\mathrm{A,FH}} + 2\theta. \tag{7}$$

The delay of one cycle with successful decoding $d_{\mathrm{B1,s}}$:

$$d_{\mathrm{B1,s}} = T_{\mathrm{D,R}} + \frac{r}{c} + T_{\mathrm{D,FH}} + \theta. \tag{8}$$

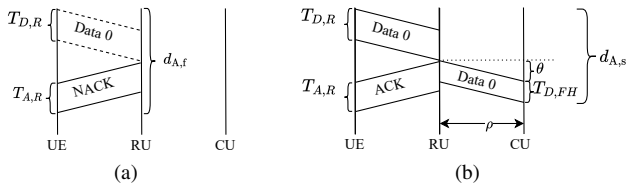

Fig. 3: 1 cycle delay (a) failure and (b) success case (Architecture A).

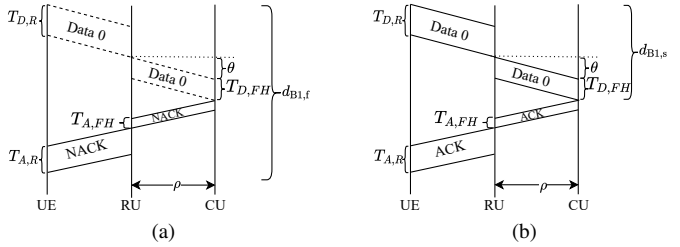

Fig. 4: 1 cycle delay (a) failure and (b) success case (Architecture B1).

Architecture B1 is split D from [6]. In this split, each decoded packet is transmitted over the fronthaul. As a response, an ACK or NACK can be transmitted over the fronthaul. The transmission on this link depends on the packet size (L_{P}) or the ACK/NACK size (L_{A}) and the corresponding headers length. For data packet transmission we have

$$T_{\mathrm{D,FH}} = \frac{L_{\mathrm{P}} + L_{\mathrm{H_{DLL}}} + L_{\mathrm{H_{ecpri}}} + L_{\mathrm{H_{TN}}}}{C_{\mathrm{FH}}}, \tag{9}$$

where $L_{\mathrm{H_{DLL}}}$ is the data link layer headers size. For an ACK/NACK transmission:

$$T_{\mathrm{A,FH}} = \frac{L_{\mathrm{A}} + L_{\mathrm{H_{DLL}}} + L_{\mathrm{H_{ecpri}}} + L_{\mathrm{H_{TN}}}}{C_{\mathrm{FH}}}. \tag{10}$$

The total propagation and transmission delay caused by $l-1$ failed transmissions and one successful, for architecture B1 is:

$$d_{\mathrm{B1}} = (l-1)d_{\mathrm{B1,f}} + d_{\mathrm{B1,s}}. \tag{11}$$

C. Architecture B2

Architecture B2 is split E from [6]. The radio signal received on the RU is sampled and quantized. The source rate on the fronthaul is calculated as [10]:

$$\mathrm{R_{FH,B2}} = f_{\mathrm{s}} \times N_{\mathrm{IQ}} \times 2 \times F_{\mathrm{os}} \times N_{\mathrm{a}} \times \eta, \tag{12}$$

where f_{s} is the sampling frequency, N_{IQ} the number of I and Q bits (multiplied by 2 to cover both I and Q bits), F_{os} the oversampling factor, N_{a} the number of antennas, and η the CPRI forward error correction (FEC) code rate. Each symbol duration, the samples are transmitted over the fronthaul in one eCPRI frame. We define $T_{\mathrm{S,FH}}$ the transmission duration of one sampled and quantized orthogonal frequency division multiplexing (OFDM) symbol over the fronthaul:

$$T_{\mathrm{S,FH}} = T_{\mathrm{symb}} + T_{\mathrm{CP}} + \frac{\mathrm{R_{FH,B2}}(T_{\mathrm{symb}} + T_{\mathrm{CP}}) + L_{\mathrm{H_{ecpri}}} + L_{\mathrm{H_{TN}}}}{C_{\mathrm{FH}}}, \tag{13}$$

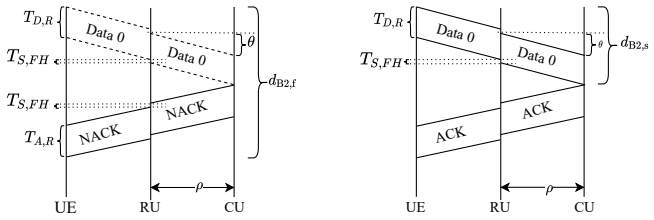

Fig. 5: 1 cycle delay (a) failure and (b) success case (Architecture B2).

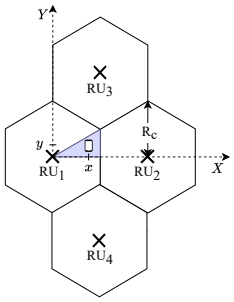

Fig. 6: Study zone (4BSs).

where T_{symb} is one symbol duration and T_{CP} the cyclic prefix duration. The cycles duration are represented in Fig. 5:

$$d_{\text{B2,f}} = T_{\text{D,R}} + T_{\text{A,R}} + 2\frac{r}{c} + 2T_{\text{S,FH}} + 2\theta. \tag{14}$$

$$d_{\text{B2,s}} = T_{\text{D,R}} + \frac{r}{c} + T_{\text{S,FH}} + \theta. \tag{15}$$

The total delay caused by $l-1$ failed transmissions and a successful one, for architecture B2 is :

$$d_{\text{B2}} = (l-1)d_{\text{B2,f}} + d_{\text{B2,s}}. \tag{16}$$

V. Analytic formulation

The symmetry of the hexagonal shape allows us to restrict our study to the highlighted area in Fig. 6. We consider that a UE is in the position (x_i, y_i) relative to an RU_i. We take (x_i, y_i) as a function of (x, y) which represent the coordinates relative to RU_1 coexisting in the same cell with the UE. That means that $(x_1, y_1) = (x, y)$.

We are interested in finding the distribution of the number of transmissions which is an r.v. affecting delay. More than k transmissions are needed if the kth transmission is erroneous. From reference [11], we take the expression of the probability of having a bad kth transmission as a function of the average SNR, $\overline{\gamma}$:

$$\begin{aligned}\mathbb{P}(l > k/\overline{\gamma}) &= \Gamma_l(k, G) \\ &+ e^{-G}\sum_{i=0}^{k-1}\frac{(G)^i}{i!}\frac{\Gamma(\frac{1}{g\overline{\gamma}})}{(g\overline{\gamma})^{k-i+1}\Gamma(\frac{1}{g\overline{\gamma}}+k-i+1)},\end{aligned} \tag{17}$$

where $G = \frac{\gamma_M}{\overline{\gamma}}$, 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. This probability was determined using HARQ-CC and the error model shown in (2).

The average SNR, experienced by RU_i, depends on the random position of the UE in the cell and on the random shadowing effect:

$$\overline{\gamma}(x_i, y_i, \xi_c, \xi_s) = \frac{P_\text{t}}{N}\left(\frac{r_0}{\sqrt{x_i^2+y_i^2}}\right)^\alpha e^{\xi_c+\xi_s}, \tag{18}$$

where N represents noise and interference and is considered constant. The probability mass function (PMF) of the number of transmissions is then computed using:

$$\mathbb{P}(l = k/\overline{\gamma}) = \mathbb{P}(l > k-1/\overline{\gamma}) - \mathbb{P}(l > k/\overline{\gamma}). \tag{19}$$

For simplicity, we let $A_i = \mathbb{P}(l > k/\overline{\gamma}(x_i, y_i, \xi_c, \xi_s))$ which is given by (17). We use A_i in the remaining equations of the article.

A. Architecture A, Nearest BS

In architecture A, a UE is served by one and only one RU. In this section, we assume that each UE is served by the nearest RU. To get the total probability of having a bad reception at the kth try, we average (17) over the area in question and over the different shadowing values as well:

$$\mathbb{P}(l > k) = q\int_0^{\frac{\sqrt{3}}{2}R_\text{c}}\int_0^{\frac{x}{\sqrt{3}}}\int_{-\infty}^{+\infty} A\frac{1}{\sqrt{\sigma_c^2+\sigma_s^2}}e^{-\frac{u^2}{2(\sigma_c^2+\sigma_s^2)}}\mathrm{d}u\mathrm{d}y\mathrm{d}x, \tag{20}$$

where $q = \frac{8}{\sqrt{2\pi}\sqrt{3}R_\text{c}^2}$, $A = \mathbb{P}\left(l > k/\overline{\gamma}(x, y, u)\right)$, and u represents $\xi_c + \xi_s$.

B. Architecture A, Best BS

When shadowing is considered, the nearest BS is not always the best serving BS. In this section, the UE is assumed to be connected to the best receiving RU among the M nearest RUs. We let RU_i be the best receiving RU with the best received SNR. Assuming constant noise, the best signal can be determined by the highest received power:

$$\mathbb{P}(\text{UE connected to RU}_i) = \mathbb{P}(P_{r,i} > P_{r,j}\ \forall j \neq i), \tag{21}$$

where $1 \leq i \leq M$, $1 \leq j \leq M$ and $P_{r,i}$ is the power received by RU_i from the UE when all UEs use the same transmission power. The received powers are independent because the M RUs are not co-located. Thus, we get the probability of being connected to RU_i:

$$\begin{aligned}&\mathbb{P}(\text{UE connected to RU}_i) = \\ &\int_{-\infty}^{+\infty}\prod_{j\neq i}\mathbb{P}\left(\xi_{sj} < u + \alpha\ln\left(\frac{\sqrt{x_j^2+y_j^2}}{\sqrt{x_i^2+y_i^2}}\right)\right)\frac{e^{-\frac{u^2}{2\sigma_s^2}}}{\sigma_s\sqrt{2\pi}}\mathrm{d}u.\end{aligned} \tag{22}$$

Let $\mathbb{P}_i = \prod_{j\neq i}\mathbb{P}\left(\xi_{sj} < \xi_{si} + \alpha\ln\frac{\sqrt{x_j^2+y_j^2}}{\sqrt{x_i^2+y_i^2}}\right)$. The probability in the previous product represents the cumulative distribution function (CDF) of ξ_{sj} and is calculated as $\frac{1}{2}\times\left(1+\text{erf}\left(\frac{w}{\sigma_s\sqrt{2}}\right)\right)$ where $w = u + \alpha\ln\frac{\sqrt{x_j^2+y_j^2}}{\sqrt{x_i^2+y_i^2}}$. The total

distribution of the number of transmissions, while being connected to the best BS, is given by:

$$\mathbb{P}(l>k)=\frac{q}{\sqrt{2\pi}}\int_0^{\frac{\sqrt{3}}{2}R_c}\int_0^{\frac{x}{\sqrt{3}}}\sum_{i=1}^{M}\left[\int_{-\infty}^{+\infty}\left(\mathbb{P}_i\int_{-\infty}^{+\infty}\frac{A_i}{\sigma_c}e^{-\frac{u^2}{2\sigma_c^2}}\mathrm{d}u\right)\times\frac{1}{\sigma_s}e^{-\frac{v^2}{2\sigma_s^2}}\mathrm{d}v\right]\mathrm{d}y\mathrm{d}x. \tag{23}$$

C. Architecture B1

For architecture B1, we consider that M RUs are receiving the signal transmitted by the UE. We consider that when all the M RUs experience a bad reception, an error is detected. The different receptions are independent and thus the probability of error is a product of errors occurring on each RU, which is then averaged over the surface in question:

$$\mathbb{P}(l>k)=\frac{q}{\sqrt{2\pi}}\int_0^{\frac{\sqrt{3}}{2}R_c}\int_0^{\frac{x}{\sqrt{3}}}\int_{-\infty}^{+\infty}\left(\prod_{i=1}^{M}\int_{-\infty}^{+\infty}\frac{A_i}{\sigma_s}e^{-\frac{u^2}{2\sigma_s^2}}\mathrm{d}u\right)\times\frac{1}{\sigma_c}e^{-\frac{v^2}{2\sigma_c^2}}\mathrm{d}v\mathrm{d}y\mathrm{d}x. \tag{24}$$

D. Architecture B2, general case

With architecture B2, M RUs receive the signal transmitted by the UE. The different receptions are combined by the MRC technique. During each transmission, the CU processes the sum of M signals received on M RUs. An error is identified when the sum can not be decoded correctly. The total SNR experienced at the CU during the kth transmission is:

$$\gamma_{S,k}=\sum_{l=1}^{M}\gamma_{l,k}, \tag{25}$$

where $\gamma_{l,k}$ is the SNR perceived at RU$_l$ during the kth transmission.

Since every reception consists now of the sum of M SNRs, (17) is not valid anymore. We need to derive the probability of error at the kth transmission. However, repeating the calculation steps done in [11] is unfeasible for this case. We thus adopt a pure simulation approach for architecture B2 in the general case. However, for a specific position of the terminal, a computation is possible as explained in the next paragraph.

E. Architecture B2, particular case

In this section, we develop the probability of having an error at the kth transmission for the MRC case with particular considerations that are valid only for the analytic calculation. We consider only two receiving RUs: $M=2$. We also consider $\overline{\gamma}_1=\overline{\gamma}_2=\overline{\gamma}$, with $\overline{\gamma}_l$ being the mean of the exponential r.v. $\gamma_{l,k}$. Thus, $\gamma_{S,k}=\gamma_{1,k}+\gamma_{2,k}$ is an Erlang r.v. with the following distribution:

$$f_\gamma(\gamma_{S,k})=\left(\frac{1}{\overline{\gamma}}\right)^2\gamma_{S,k}e^{-\frac{\gamma_{S,k}}{\overline{\gamma}}}. \tag{26}$$

Note that a simulation approach is used for other considerations ($M>2$ for example). We again use HARQ-CC. During the kth transmission, the SNR used to determine the PER in (2) is:

$$\gamma_{T,k}=\sum_{i=1}^{k}\gamma_{S,i}. \tag{27}$$

The PER during the kth transmission is therefore $h(\gamma_{T,k})$. The probability of having more than k transmissions results from having errors during all of the first k transmissions. So, it depends on all the previous SNRs (all γ_i with $i\leq k$). We consider successive packet transmissions, so the SNR is independent and identically distributed (i.i.d) for different transmissions. Therefore, the probability of error during the kth transmission is the following:

$$\mathbb{P}(l>k)=\int_0^{\infty}\!\!\ldots\int_0^{\infty}\prod_{i=1}^{k}h(\gamma_{T,i})f_\gamma(\gamma_{S,1})\ldots f_\gamma(\gamma_{S,k})\mathrm{d}\gamma_{S,1}\ldots\mathrm{d}\gamma_{S,k}. \tag{28}$$

Similarly to [11], we split each integral into two integrals, resulting in:

$$\mathbb{P}(l>k)=B_k+\sum_{l=1}^{k-1}C_{k,l}+D_k. \tag{29}$$

After several calculation steps, omitted for the sake of brevity and which can be found in [12], we get:

$$B_k=1-e^{-\frac{\gamma_M}{\overline{\gamma}}}\sum_{n=0}^{2k-1}\left(\frac{\gamma_M}{\overline{\gamma}}\right)^n\frac{1}{n!}, \tag{30}$$

$$C_{k,l}=e^{-\frac{\gamma_M}{\overline{\gamma}}}\frac{1}{(2l+1)!}\left(\frac{\gamma_M}{\overline{\gamma}}\right)^{2l}\prod_{j=l+1}^{k}\frac{1}{(1+g\overline{\gamma}(k+1-j))^2}\times\left[\gamma_M\left(\frac{1}{\overline{\gamma}}+g(k-l)\right)+2l+1\right], \tag{31}$$

$$D_k=e^{-\frac{\gamma_M}{\overline{\gamma}}}\prod_{j=1}^{k}\frac{1}{(1+g\overline{\gamma}(k+1-j))^2}\left[1+\gamma_M\left(\frac{1}{\overline{\gamma}}+gk\right)\right]. \tag{32}$$

Finally, by substituting B_k, $C_{k,l}$, and D_k in (29) by (30), (31), and (32) respectively, we get $\mathbb{P}(l>k)$ for the MRC case:

$$\mathbb{P}(l>k)=1-e^{-G}\sum_{i=0}^{2k-1}\frac{G^i}{i!}+\sum_{l=0}^{k-1}\frac{e^{-G}G^{2l}T_{k,l}}{(2l+1)!}\prod_{j=1}^{k-l}\frac{1}{(1+\overline{\gamma}gj)^2}, \tag{33}$$

where $T_{k,l}=\gamma_M\left(\frac{1}{\overline{\gamma}}+g(k-l)\right)+2l+1$.

VI. Results and discussions

We carried out both simulations and computations to check whether they gave the same results. In one simulation, 100 000 users were uniformly distributed in the shadowed area of Fig. 6. For each user, fading and shadowing were randomly generated. The fading changed with each transmission while the shadowing was considered the same for the same UE for

TABLE I: Parameters values.

Symbol	Parameter	Values
α	Path-loss exponent	3.38
δ	Correlation coefficient	0.5
η	CPRI FEC code rate	$\frac{10}{8}$
ρ (Km)	CU-RUs distance	3.5
σ_{dB} (dB)	Shadowing's standard deviation	5
a [9]	Parameter depending on the MCS	274.7
c (m/s)	Light velocity	3×10^8
C_{FH} (Gbps)	Fronthaul maximum capacity	100
F_{os}	Over sampling factor	2
f_{s} (Msamples/s)	Sampling frequency	153.6
g [9]	Parameter depending on the MCS	7.993
L_{A} (Bits)	ACK/NACK length	1
$L_{\text{H}_{\text{DLL}}}$(Bytes)	Data link layer header (SDAP, PDCP, RLC, MAC)	11
$L_{\text{H}_{\text{ecpri}}}$(Bytes)	eCPRI header	4
$L_{\text{H}_{\text{TN}}}$(Bytes)	Transport-network header (UDP, IP, ethernet)	62
L_{P} (Bytes)	Data packet length	32 [13]
N (dBm)	Noise power	-116
N_{a}	Number of antennas	1
N_{IQ} (bits)	Number of I and Q bits	16
P_{t} (dBm)	UE's transmission power	23
r_0 (m)	Reference distance	0.2
R_{c} (km)	Cell radius	2.2
$T_{\text{A,R}}$ (ms)	ACK/NACK transmission duration over the radio interface	0.25 [a]
$T_{\text{D,R}}$ (ms)	Data transmission duration over the radio interface	0.25 [a]
$T_{\text{CP}}(\mu s)$	Cyclic prefix duration	1.17 [a]
$T_{\text{symb}}(\mu s)$	Symbol duration	16.67 [a]

[a] Numerology 2 of the 5G new radio (NR) [14].

TABLE II: PMF of the number of transmissions for architectures A and B1 with simulations confidence margins.

Number of transmissions (k)	PMF (Analytic)	PMF (Simulations average)	95% confidence margin
Nearest BS-A			
1	0.8805	0.8803	$[0.8774, 0.8832]$
2	0.0925	0.0924	$[0.0900, 0.0950]$
3	0.0183	0.0183	$[0.0172, 0.0194]$
Best BS(2 nearest)-A			
1	0.8976	0.8977	$[0.8958, 0.8995]$
2	0.0844	0.0843	$[0.0826, 0.0860]$
3	0.0135	0.0135	$[0.0129, 0.0142]$
2 BSs-B1			
1	0.9420	0.9420	$[0.9406, 0.9434]$
2	0.0497	0.0497	$[0.0484, 0.0510]$
3	0.0063	0.0062	$[0.0058, 0.0067]$

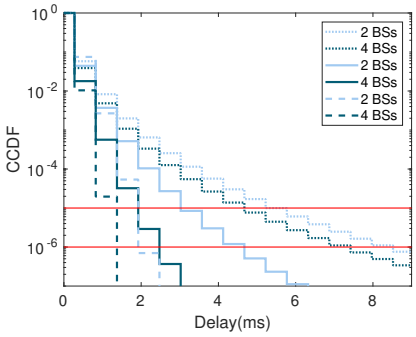

Fig. 7: Delay CCDF for architecture B1 with $\sigma_{\text{dB}} = 5$ dB correlated shadowing (dotted lines), $\sigma_{\text{dB}} = 5$ dB decorrelated shadowing (continuous lines) and $\sigma_{\text{dB}} = 0$ dB (dashed lines).

all the transmissions. The PER is given by (2). The simulation was iterated 1000 times. The confidence margin is evaluated at 95%. For the results, numerical integration method is used to compute the probabilities in equations (20), (23), and (24). The simulation and calculation parameters are summarized in Table I. We consider an upper bound for the propagation delay over the radio interface: $\frac{r}{c} = \frac{R_{\text{c}}}{c}$.

Table II shows the similarity of the results between the simulation and our mathematical computation for the PMF of the number of transmissions. It compares the PMF for three cases: receiving from the nearest BS, receiving from the best BS among the two nearest BSs (architecture A), and receiving from the two nearest BSs (architecture B1). We can see the improvement when the best BS is selected. In such cases, fewer transmissions are needed to get a correct packet. An additional improvement is observed when macro-diversity is used with architecture B1.

Knowing the distribution of the number of transmissions, we can get the distribution of the delay. The complementary cumulative distribution function (CCDF) of the delay in figures 7 and 9 represents the probability of not receiving a good packet within a certain amount of time. In other words, this CCDF represents the PER. The importance of macro-diversity is highlighted in Fig. 7. When shadowing is not considered, macro-diversity does not improve as much as when a decorrelated shadowing ($\delta = 0$) is considered. For instance, for a PER of 10^{-6}, going from from 2 to 4 receiving BSs reduces 0.55 ms in terms of latency with $\sigma = 0$ dB. Whereas, with 5 dB decorrelated shadowing deviation, we have a reduction of 2.2 ms. When the shadowing is correlated, moving from 2 to 4 receiving BSs does not improve a lot. The improvement is quietly similar to the non-shadowed case (see 10^{-5} PER also). For a PER of 10^{-6}, this improvement is a reduction of 1.1 ms. This is due to the fact that when the shadowing is correlated, the performance of the nearest BS approaches the performance of the best one. Nevertheless, it is worth mentioning that spatial diversity, with architecture B1, improves compared to only receiving from the nearest BS, when the shadowing is correlated (Fig.9). But, a high diversity order is not required for an additional improvement.

We get similar analytic and simulation results for the MRC with architecture B2 as shown in Table III. Fig. 8 shows that receiving from two BSs, whether using the first combining technique or the second one, improves the average number of transmissions. Lower average number of transmissions produces lower average latency. For $\overline{\gamma} = \overline{\gamma}_1 = \overline{\gamma}_2 = -2$ dB, we have an average delay of 0.915 ms for the nearest BS with architecture A, compared to 0.6350 ms and 0.4973 ms with architectures B1 and B2, respectively. So, we notice that MRC outperforms the combining technique used with architecture B1. Nevertheless, when both channels, between the UE and both RUs, are good (high $\overline{\gamma}_1$ and $\overline{\gamma}_2$), B1 and B2 have almost

TABLE III: PMF of the number of transmissions and 95% simulation confidence margin for the MRC technique (architecture B2, particular case).

Number of transmissions (k)	**PMF (Analytic)**	**PMF (Simulations average)**	**95% confidence margin**
$\overline{\gamma} = -3$ dB			
1	0.5124	0.5124	$[0.5096, 0.5154]$
2	0.3997	0.4008	$[0.3969, 0.4024]$
3	0.0802	0.0794	$[0.0787, 0.0817]$
$\overline{\gamma} = 0$ dB			
1	0.7985	0.7970	$[0.7961, 0.8008]$
2	0.1907	0.1921	$[0.1885, 0.1930]$
3	0.0105	0.0107	$[0.0100, 0.0110]$
$\overline{\gamma} = 3$ dB			
1	0.9337	0.9338	$[0.9323, 0.9352]$
2	0.0653	0.0653	$[0.0639, 0.0667]$

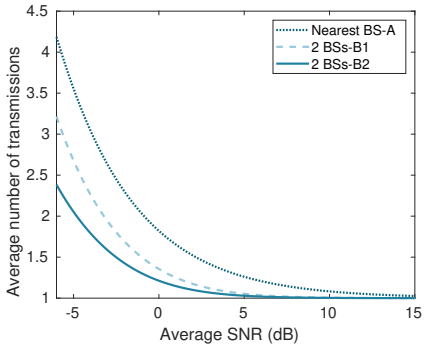

Fig. 8: Average number of transmissions as a function of $\overline{\gamma}$ when the average SNR is the same on both sites: $\overline{\gamma}_1 = \overline{\gamma}_2 = \overline{\gamma}$.

the same performance.

Fig. 9 compares the three architectures with all the cases under study. The results of architecture B2 shown here are the simulation results. We can see the improvement produced when receiving from the best BS, compared to the nearest one. An additional improvement is noticed when receiving form 4 BSs with architecture B1. The improvement rises when using MRC with architecture B2. In fact, summing the signals before decoding increases the chances of good decoding, i.e. high reliability, and decreases the number of re-transmissions, i.e. low latency. The difference appears to be considerable for low PER. If we take a PER of 10^{-2}, the difference between the delay of architecture B1 and architecture B2 is 0.06 ms, with a shorter delay for architecture B1. On the other hand, for an ultra reliability of 10^{-6}, this difference increases to approximately 3.59 ms, with a shorter delay using architecture B2. So, for error-tolerant applications, architecture B1 is sufficient. However, for URLLC applications, architecture B2 with MRC is better.

VII. Conclusion

In this paper, we studied the impact of architecture and macro-diversity on reliability and latency with three different functional splits. The comparison has been made through analytic calculations and simulations. It was shown that receiving from the best BS induces less latency and higher reliability than receiving from the nearest BS. Macro-diversity with architectures B1 and B2 had a better impact on reliability and latency compared to a single reception with architecture A. The main finding of this article is the importance of the MRC when both high reliability and low latency are required. For future work, we propose studying the load on the fronthaul, and then to choose the optimal FS in terms of reliability, latency, and fronthaul charge.

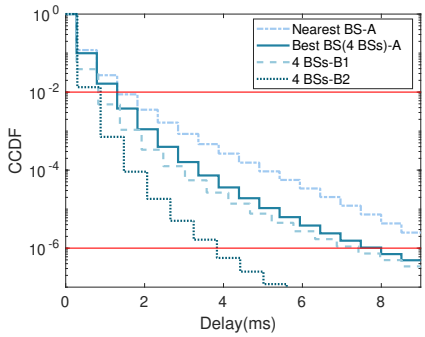

Fig. 9: Delay CCDF for architectures A (2 cases), B1, and B2.

References

[1] J. Sachs, G. Wikstrom, T. Dudda, R. Baldemair, and K. Kittichokechai, "5G radio network design for ultra-reliable low-latency communication," *IEEE Network*, vol. 32, no. 2, pp. 24–31, 2018.

[2] 3GPP, "5G; Service requirements for next generation new services and markets," 3rd Generation Partnership Project, TS 22.261, July 2017.

[3] N. Strodthoff, B. Göktepe, T. Schierl, C. Hellge, and W. Samek, "Enhanced machine learning techniques for early HARQ feedback prediction in 5G," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 11, pp. 2573–2587, 2019.

[4] 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.

[5] G. Mountaser, M. Condoluci, T. Mahmoodi, M. Dohler, and I. Mings, "Cloud-RAN in support of URLLC," in *2017 IEEE GC Workshops*, 2017.

[6] "eCPRI Specification V2.0," Interface Specification, May 2019.

[7] T. Alhajj and X. Lagrange, "Reliability and low latency: impact of the architecture," in *IEEE ISCC*, 2020.

[8] 3GPP, "Evolved universal terrestrial radio access (E-UTRA); radio frequency (RF) system scenarios," 3rd Generation Partnership Project, TR 36.942.

[9] 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 on*, vol. 3, pp. 1746 – 1755, October 2004.

[10] J. Duan, X. Lagrange, and F. Guilloud, "Performance analysis of several functional splits in C-RAN," in *IEEE 83rd VTC Spring*, 2016.

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

[12] T. Alhajj and X. Lagrange, "Computation of the probability of error for HARQ-CC with macro-diversity based on MRC." IMT ATLANTIQUE, Research Report, Sep. 2020. [Online]. Available: https://hal-imt.archives-ouvertes.fr/hal-02949519

[13] 3GPP, "5G; Study on scenarios and requirements for next generation access technologies," 3rd Generation Partnership Project, TR 38.913, May 2017.

[14] 3GPP, "5G NR physical channels and modulation," 3rd Generation Partnership Project, TS 38.211, July 2018.