Abstract—Cloud-Radio Access Network (C-RAN), one of the prominent architectures for 5G cellular systems, is characterized by a hierarchical structure in which the baseband processing functionalities of remote radio heads (RRHs) are implemented by means of cloud computing at a Central Unit (CU). A key limitation of C-RANs is given by the capacity constraints of the fronthaul links connecting RRHs to the CU. In this letter, the impact of this architectural constraint is investigated for the fundamental functions of random access and active User Equipment (UE) identification in the presence of a potentially massive number of UEs. In particular, two baseline algorithmic solutions are studied, namely a standard C-RAN approach based on quantize-and-forward and an alternative scheme based on detect-and-forward. Both techniques leverage Bayesian compressive sensing (CS) algorithms and are elucidated with reference to the state-of-the-art on CS. Numerical results illustrate the advantages of centralized processing and the relative merits of the two schemes as a function of the system parameters.

I. INTRODUCTION

Cloud-Radio Access Network (C-RAN) is established as one of the most prominent architectures for 5G cellular systems due to the cost reductions and performance advantages that arise from the implementation of baseband processing at a centralized cloud processor or Central Unit (CU) on behalf of multiple distributed Remote Radio Heads (RRHs). This migration is made possible by deploying links that connect the RRHs to the CU, with possible media that include fiber optic cables, DSL last-mile links or wireless mmwave channels. The capacity limitations of these links, along with the associated latency, are understood to offer the most significant challenge to the implementation of C-RANs [1].

A fundamental network function is random access, which is carried out by user equipments (UEs) when first accessing the system. Random access is attracting renewed interest due to the expected increase in the number of UEs in 5G networks, with particular reference to massive access in Internet-of-Things applications (see, e.g., [2]). One of the main goals of the random access procedure is for the network to identify the set of active UEs in order to enable resource allocation. This is typically accomplished by having the UEs transmit signature sequences which serve as UE identifiers, with possibly successive transmission rounds needed to resolve collisions due to multiple UEs selecting the same signature.

In this letter, we study user activity detection (UAD) (sometimes also referred to as active user detection, as in, e.g., [3]) for a C-RAN architecture with the aim of investigating solutions that address the mentioned fronthaul capacity limitations. We assume that the UEs employ non-orthogonal sequences so as to accommodate a potentially massive number of UEs, e.g., machine-type devices, and we do not assume any a priori knowledge of the instantaneous small-scale fading channel realizations. Under the further assumption that the number of active UEs is significantly smaller than the total number of UEs, the signal received at the RRH is sparse with respect to the set of the UE signatures. As a result, the UAD problem becomes one of sparse signal recovery, which can be studied within a compressive sensing (CS) framework (see, e.g., [4], [5], [6] and references therein). A CS-based algorithm for UAD in C-RANs was recently proposed in [3] using a Bayesian formulation under the assumption of ideal, i.e., infinite-capacity, fronthaul links. With the aim of exploring the impact of fronthaul capacity limitations, in this letter we study two schemes, namely Quantize-and-Forward (QF) and Detect-and-Forward (DF), that differ in the type of processing performed at the RRHs and at the CU. The QF scheme amounts to the standard C-RAN implementation, whereby the RRHs quantize the received samples for transmission on the fronthaul links to the CU, which performs baseband...
processing for UAD (see, e.g., [1]). The DiF scheme, instead, deviates from the classical C-RAN architecture by adopting an alternative CU-RRH functional split, in the sense of, e.g., [2], in which part of the baseband processing is carried out at the RRHs. In particular, with DiF, each of the RRHs performs local UAD, and forwards quantized soft information associated with the local decision in the form of log-likelihood ratios (LLRs) to the CU. The two schemes are compared via numerical results in terms of the trade-off between the fraction of correctly detected active UEs and the fraction of incorrectly detected inactive UEs.

Notation: Uppercase/lowercase boldface letters denote matrices/vectors. Random quantities are represented with standard fonts, while italic is used for deterministic quantities. The superscript $H$ stands for Hermitian transposition and $\otimes$ stands for Kronecker matrix multiplication. $\mathcal{N}(\mu, \sigma^2)$ and $\mathcal{CN}(\mu, \sigma^2)$ denote real and circularly symmetric, respectively, Gaussian random variable with expectation $\mu$ and variance $\sigma^2$. The notation $[X_1, \ldots, X_N]$ for matrices $X_1, \ldots, X_N$ of suitable sizes represents a matrix that stacks $X_1, \ldots, X_N$ vertically, while $[X_1; \ldots; X_N]$ stacks $X_1, \ldots, X_N$ horizontally.

II. SYSTEM MODEL

We consider a slotted random access system with $N$ UEs. The user activity (random) variable $\lambda_n \in \{0, 1\}$ equals 1 if UE $n$ is active in the given block, which happens with probability $p$, and 0 otherwise. When active, UE $n$ transmits over $M$, in general, complex symbols of the time-frequency grid the identification signature 

$$s_n = [s_{n,1}, \ldots, s_{n,M}].$$

(1)

The UE signatures are subject to the energy constraint $E[\|s_n\|^2] = E_s$. We assume that the time-frequency grids of different users are aligned, which requires time synchronization.1 Moreover, focusing on a scenario with a potentially massive number $N$ of UEs, the signatures are assumed to be nonorthogonal. Assuming a block-fading model with coherence time-frequency span no smaller than that occupied by the signatures’ transmission, the signal $w_r = [w_1, \ldots, w_M]$ received at RRH $r$, $r = 1, \ldots, R$, reads

$$w_r = \sum_{n=1}^{N} \lambda_n \gamma_{n,r} h_{n,r} s_n + v_r,$$

(2)

where $\gamma_{n,r} \in \mathbb{R}$ and $h_{n,r} \in \mathbb{C}$ are the large- and small-scale fading coefficients respectively, for the link between the $n$-th UE and the $r$-th RRH, and $v_r$ is an additive noise vector. We further assume that the coefficients $h_{n,r}$ are independent identically distributed (i.i.d.) $\mathcal{CN}(0, 1)$ and the elements of the noise vector $v_r$ are i.i.d. $\mathcal{CN}(0, \sigma^2_v)$. The large scale fading coefficients $\gamma_{n,r}$ are assumed to be known to the RRHs and to the CU as in, e.g., [3], in contrast to the unknown small scale fading coefficients $h_{n,r}$. We assume that the UEs, RRHs and the CU know the small-scale fading statistics and the probability of UE activation $p$.

The fronthaul capacity limitations are expressed in terms of the number of bits per received complex sample, $b_r$ available

1Our model may be extended to include frame asynchronicity (with symbol-level synchronization intact) by using cyclic-extended signature waveforms based, for example, on Gabor frames or Kerdock codes [5].

for transmission on the fronthaul link between RRH $r$ and the CU. The system model is illustrated in Fig. 1. To elaborate further, we rewrite the received signal (2) as

$$w_r = S \Gamma_r A h_r + v_r,$$

(3)

where the columns of $S \in \mathbb{C}^{M \times N}$ represent the UEs’ signatures; $\Gamma_r \in \mathbb{R}^{N \times N}$ is a diagonal matrix with $\gamma_{n,r}$, $n = 1, \ldots, N$, on the main diagonal; $h_r \in \mathbb{R}^{N \times 1}$ is a vector of the small-scale fading coefficients; and $A$ is a diagonal matrix with vector $\lambda = [\lambda_1, \ldots, \lambda_N]$ on the main diagonal. We can further simplify (3) as

$$w_r = A_r x_r + v_r,$$

(4)

with the definitions $A_r = S \Gamma_r$ and $x_r = A h_r$. Note that the columns of $A_r$ are the signatures of the UEs scaled with the corresponding large scale fading coefficients of the links between the UEs and $r$-th RRH, which are assumed to be known to the RRH, while $x_r$ depends on the unknown user activity and small-scale fading variables.

The model (4) can be interpreted within a Bayesian formulation of the CS problem (see, e.g., [8]). In fact, the unknown $x_r \in \mathbb{C}^{N \times 1}$ is a Bernoulli-Gaussian random vector, with entries $x_{n,r}$, $n = 1, \ldots, N$ being equal to the channel $h_{n,r} \sim \mathcal{CN}(0,1)$ with probability $p$, accounting for the case of an active UE $n$ ($\lambda_{n} = 1$), or else equal to zero with probability $1-p$ when UE $n$ is inactive ($\lambda_{n} = 0$). When $p$ is small, and in the absence of fronthaul capacity limitations, UAD hence translates into estimating the support of a sparse i.i.d. Bernoulli-Gaussian vector from a Multiple Measurement Vector (MMV) model (see, e.g., [8]), as investigated in [3].

III. FRONTHAUL AND UAD PROCESSING

In this section, we discuss two baseline schemes that account for fronthaul capacity limitations, namely QF and DiF.

A. Quantize-and-forward (QF)

With QF, each RRH $r$ quantizes the measurement $w_r$ in (4) with $b_r$ bits per sample, and forwards the quantized samples to the CU. The signal received by the CU on the $r$-th fronthaul can hence be written as

$$y_r = Q_r(w_r) = Q_r(A x_r + v_r),$$

(5)

where $Q_r$ is a quantization function, applied element-wise to the entries of the argument vector, with resolution $b_r$ bits. We assume here that the function $Q_r$ amounts to two scalar uniform quantizers with $2^{b_r/2}$ levels applied to the real and imaginary parts of the entries of $w_r$. The dynamic range is selected so as to capture three standard deviations for both positive and negative values of each real component.

The overall signal $y = [y_1, \ldots, y_R]$ retrieved by the CU from the fronthaul links can then be expressed in a compact fashion by defining $\Gamma^{(n)}$ as the diagonal matrix with the vector $[\gamma_{1,n}, \ldots, \gamma_{N,n}]$ of long-term fading coefficients on the main diagonal along with the unknown vector $x^{(n)} = [x_{n,1}, \ldots, x_{n,R}]$. In particular, we can write

$$y = Q(w) = Q(A x + v),$$

(6)

where we have $A = [\Gamma^{(1)} \otimes s_1; \ldots; \Gamma^{(N)} \otimes s_N]$, $x = [x^{(1)}; \ldots; x^{(N)}]$, $w = [w_1, \ldots, w_R]$ and $Q$ is to be under-
stood as being the same as \( Q_r \) whenever it is applied to a component of \( w \) coming from \( w_r \).

The CU performs UAD based on the received signal (6), by implementing a CS reconstruction algorithm that aims at estimating the support of the Bernoulli-Gaussian vector \( x \) from the linear mixture \( z + v \), with \( z = Ax \), as observed after the application of the sample-by-sample non-linearity \( Q \).

In particular, the unknown vector \( x \) in (6) is characterized by group sparsity, since each subvector \( x^{(n)} \) equals an all-zero vector if UE \( n \) is not active, i.e., if \( \lambda_n = 0 \), and is generally non-zero when \( \lambda_n = 1 \). For the purpose of CS reconstruction, here we adopt the Hybrid Generalized Approximate Message Passing (H-GAMP) method developed in [9], which extends over the Generalized AMP scheme (GAMP) [10] to accommodate group sparsity. This is briefly elaborated on next.

H-GAMP, as GAMP, is based on a quadratic approximation of the sum-product message passing scheme and operates by exchanging messages on the factor graph that describes the joint distribution \( p_{\lambda,x,y}(\lambda,x,y) \). More precisely, since the H-GAMP algorithm operates on real-valued variables, we first redefine the signal model (6) as follows: (i) each entry of the vectors \( x \) and \( y \) is substituted by two real entries corresponding to its real and imaginary parts; (ii) each of the entries \( a_{ij} \) of matrix \( A \) is substituted by the submatrix

\[
\begin{pmatrix}
\Re(a_{ij}) & -\Im(a_{ij}) \\
\Im(a_{ij}) & \Re(a_{ij})
\end{pmatrix}
\]

Note that each subvector \( x^{(n)} \) is now of size \( 2R \), instead of \( R \), but the group sparsity properties of the vector \( x \) are preserved.

Denoting by \( x \) and \( y \) the real-valued vectors introduced above, the joint distribution of the \( \lambda, x \) and \( y \) for which the H-GAMP algorithm operates factors as

\[
p_{\lambda,x,y}(\lambda,x,y) = \prod_{n=1}^{N} p_\lambda(\lambda_n) \prod_{j=1}^{2Rn} p_{\xi(x_j)}(\xi(x_j)) \prod_{i=1}^{2RM} p_{\mu(z_i)}(z_i),
\]

where \( \xi(j) \) denotes the index of the UE that corresponds to entry \( x_j \); and \( z_i = a_i^T x + v_i \), with \( a_i \) being the \( i \)-th row of \( A \) and \( v_i \sim CN(0,\sigma_i^2/2) \). In (7), we have \( p_\lambda(\lambda) = p_\lambda(1-p)^{1-\lambda} \); \( p_{\xi(x_j)}(\xi(x_j)) \) amounts to the \( CN(0,1) \) probability density function if \( \lambda(x_j) = 1 \), and to a Kronecker delta function centred at \( x_j = 0 \) if \( \lambda(x_j) = 0 \); and

\[
p_{\mu(z_i)}(z_i) = \int_{Q^{-1}(z_i)} \phi(u; z_i, \sigma^2) \, du,
\]

where \( Q^{-1} \) is the inverse of the component of the quantization function \( Q \) that applies to the entry \( y_i \) and \( \phi(u; z_i, \sigma^2) \) denotes the probability density function of a Gaussian vector with mean \( z_i \) and variance \( \sigma^2 \). Note that the factorization (7) uses the fact that, for a given UE activity pattern \( \lambda \), the small-scale channel coefficients in \( x \) are independent. The H-GAMP algorithm, which is detailed in [9] and can be directly applied to (7), outputs an approximation of the posterior distributions \( p_{\lambda|y}(\lambda_n|y) \) for all UE \( n = 1,\ldots,N \). From these probabilities, the log-likelihood ratio (LLR) \( l_n \), associated with the belief that UE \( n \) is active is computed as \( l_n = \log \frac{p_{\lambda|y}(\lambda_n=1|y)}{p_{\lambda|y}(\lambda_n=0|y)} \) . Based on the LLR \( l_n \), the CU estimates the user activity variable as \( \lambda_n = 1 \) if \( l_n \geq l_{th} \) for some threshold \( l_{th} \) and \( \lambda_n = 0 \) otherwise.

**Remark 1:** In the presence of a packetized fronthaul transmission, e.g., via Ethernet, instead of a rate constraint on the fronthaul link, it is relevant to consider a constraint on the overall number of bits \( B = Mb \) that the RRH can communicate to the CU. In this case, it is possible to trade the signature length, say \( M' \), with the number of bits per complex sample, say \( b' \), under the constraint \( B = M'b' \). In the single measurement vector (SMV) setting, this problem has been studied in the framework of recovery of sparse signals from 1-bit measurements in [11],[12], and for quantized measurements with multiple quantization levels in [13].

**B. Detect-and-forward (DtF)**

With DtF, each of the RRHs performs a local estimate of the UEs’ activity pattern \( \{\lambda_n\} \) and then forwards quantized soft information on these estimates to the CU over the capacity limited fronthaul links. Specifically, each RRH \( r \) estimates the LLR \( l_{r,n} = \log \frac{p_{\lambda|y}(\lambda_n=1|y_r)}{p_{\lambda|y}(\lambda_n=0|y_r)} \) associated with the belief of UE \( n \) active based on the observation \( w_r \) in (4). This can be done by means of the H-GAMP following the same approach discussed above. Each RRH then quantizes each LLR as

\[
\tilde{l}_{r,n} = Q_r(l_{r,n})
\]

where \( Q_r \) is a scalar quantization function applied to the (real-valued) argument. Given the fronthaul rate \( b_r \), the quantizer \( Q_r \) has a number of levels equal to \( 2^{Mb_r/N} \) since there are \( N \) LLRs to quantize with a total of \( M b_r \) bits. The dynamic range is selected based on preliminary Monte Carlo simulations to capture a 95% confidence interval.

UAD detection is finally performed at the CU by summing the LLRs obtained from all the RRHs. Specifically, for the UE \( n \), the CU computes \( \tilde{l}_n = \sum_{r=1}^{R} \tilde{l}_{r,n} \) and then applies a threshold test on the resulting LLR \( l_n \). We observe that this test is optimal, in the case of unquantized LLRs, in case the observations of the RRHs are conditionally i.i.d. given the transmitted signatures, while this is, in general, not the case here due to different large-scale fading coefficients \( \gamma_{n,r} \). Generalized rules that capture asymmetries in the quality of the observations at the RRHs can be devised but they will not be further investigated here.

**Remark 2:** In this letter, we consider scalar quantization for both QF and DtF. It should be mentioned that improved performance could be obtained by leveraging compression techniques in lieu of scalar quantization. For instance, for DtF, the RRH could first perform an estimate of the Bernoulli vector associated with the UE activity pattern and then compress it losslessly using the knowledge of the probability \( p \) (see [14] for related discussion on the compression of sparse sources).
of UEs and RRHs. The average signal-to-noise ratio (per system user) is defined as $\rho = E_s/(M\sigma^2)$. In the following,
we assume the UE signature vectors to be random, with i.i.d. circularly-symmetric complex Gaussian entries, for which the convergence of approximate message passing schemes has been studied rigorously [13, 10, 9]. We note that the described schemes are not tied to the specific choice of the signature sequences, and other constructions, e.g., based on Gabor frames or Kerdock/Reed-Mueller codes, may be used. In that case, however, the convergence of H-GAMP and approximate message passing in general has to be addressed accordingly (see [16] for related discussion).

Finally, we note that trading using a larger signature over using a more refined quantization deteriorates the performance, demonstrating the advantages of the described schemes are not tied to the specific choice of the signature sequences, and other constructions, e.g., based on Gabor frames or Kerdock/Reed-Mueller codes, may be used. In that case, however, the convergence of H-GAMP and approximate message passing in general has to be addressed accordingly (see [16] for related discussion).

We investigated the impact of fronthaul capacity limitations on the performance of QF and DiF. The comparison between the two schemes is performed here under a fixed average false alarm ratio of 0.2 (see Fig. 2 for an illustration). The signature length is $M = 128$ and we consider different values for the number $R$ of RRHs. We observe that DiF outperforms QF for stringent fronthaul capacity constraints. This is due to the fact that the performance of QF under a small fronthaul bit budget is hampered by coarseness of the fronthaul quantization (see [13] for related discussion). Instead, DiF benefits from the local UAD processing done at the RRHs to reduce the amount of information that needs to be transmitted to the CU (see also [14]). In the complementary regime in which the fronthaul capacity is sufficiently large, QF outperforms DiF. In fact, when the quantized signals are sufficiently accurate, QF benefits from the joint processing capabilities of the CU to perform UAD on the signals received by the RRHs. This is unlike DiF, in which local UAD processing prevents the CU to have direct access to the measurements of the RRHs. Finally, it is observed that the performance of UAD saturates at (relatively) moderate values of $b_r$, demonstrating that in this regime the performance is limited by the signature length $M$ rather than by the fronthaul capacity constraints.

V. Conclusions

We investigated the impact of fronthaul capacity limitations in a C-RAN architecture on the functions of random access and active UE identification in the presence of a potentially massive number UEs. In particular, we studied the performance of two baseline algorithmic solutions leveraging Bayesian compressive sensing: a C-RAN approach based on quantize-and-forward (QF) and an alternative scheme based on detect-and-forward (DiF). Numerical results illustrate the advantages of centralized processing and the relative merits of the two schemes. While here we have concentrated on the function of user activity detection, our framework could also be extended to allow for the integration of data transmission, by assigning a subset of sequences to serve as codewords for each of the system users. Future interesting work includes the analysis of the impact of more sophisticated compression techniques for fronthaul transfer on the performance of random access.
REFERENCES

[1] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, “Cloud RAN for mobile networks – a technology overview,” IEEE Communications Surveys Tutorials, vol. 17, no. 1, pp. 405–426, 2015.

[2] A. Layla, L. Alonso, and J. Alonso-Zarate, “Is the random access channel of LTE and LTE-A suitable for M2M communications? A survey of alternatives,” IEEE Communications Surveys Tutorials, vol. 16, no. 1, pp. 4–16, 2014.

[3] X. Xia, X. Rao, and V. K. N. Lau, “Active user detection and channel estimation in uplink cran systems,” in Proc. IEEE International Conference on Communications (ICC), June 2015, pp. 2727–2732.

[4] A. K. Fletcher, S. Rangan, and V. K. Goyal, “On-off random access channels: A compressed sensing framework,” arXiv preprint arXiv:0903.1022, 2009.

[5] Y. Xie, Y. Chi, L. Applebaum, and R. Calderbank, “Compressive demodulation of mutually interfering signals,” in Proc. IEEE Statistical Signal Processing Workshop (SSP), Aug 2012, pp. 592–595.

[6] G. Wunder, P. Jung, and W. Chen, “Compressive Random Access for Post-LTE Systems,” in Proc. IEEE ICC’14 Workshop on Massive Uncoordinated Access Protocols (MASSAP), Sydney, Australia, Jun. 2014.

[7] U. Dötsch, 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 Technical Journal, vol. 18, no. 1, pp. 105–128, 2013.

[8] S. Foucart and H. Rauhut, A mathematical introduction to compressive sensing, Springer, 2013.

[9] S. Rangan, A. Fletcher, V. Goyal, and P. Schniter, “Hybrid generalized approximate message passing with applications to structured sparsity,” in Proc. IEEE Int. Symp. Inform. Theory (ISIT), July 2012, pp. 1236–1240.

[10] S. Rangan, “Generalized approximate message passing for estimation with random linear mixing,” in Proc. IEEE Int. Symp. Inform. Theory (ISIT), July 2011, pp. 2168–2172.

[11] L. Jacques, J. N. Laska, P. T. Boufounos, and R. G. Baraniuk, “Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors,” IEEE Trans. Inform. Theory, vol. 59, no. 4, pp. 2082–2102, April 2013.

[12] Y. Plan and R. Vershynin, “Dimension reduction by random hyperplane tessellations,” Discrete and Computational Geometry, vol. 51, no. 2, pp. 438–461, 2014. [Online]. Available: http://dx.doi.org/10.1007/s00454-013-9561-6

[13] Y. Mruc and L. Rosasco, “q-ary compressive sensing,” arXiv:1302.5168v1, 2013.

[14] C. Weidmann and M. Vetterli, “Rate distortion behavior of sparse sources,” IEEE Trans. Inform. Theory, vol. 58, no. 8, pp. 4969–4992, Aug 2012.

[15] D. Donoho, A. Maleki, and A. Montanari, “Message-passing algorithms for compressed sensing,” Proceedings of the National Academy of Science, 2009.

[16] F. Caltagirone, L. Zdeborova, and F. Krzakala, “On convergence of approximate message passing,” in Proc. IEEE Int. Symp. Inform. Theory, June 2014, pp. 1812–1816.

[17] M. E. Davies and Y. C. Eldar, “Rank awareness in joint sparse recovery,” IEEE Trans. Inform. Theory, vol. 58, no. 2, pp. 1135–1146, Feb 2012.

[18] V. K. Goyal, A. K. Fletcher, and S. Rangan, “Compressive sampling and lossy compression,” IEEE Signal Processing Mag., vol. 25, no. 2, pp. 48–56, March 2008.