Statistical Multiplexing in Fronthaul-Constrained Massive MIMO

Jay Kant Chaudhary, Jens Bartelt and Gerhard Fettweis
Vodafone Chair Mobile Communications Systems
Technische Universität Dresden, Germany
Email: {jay_kant.chaudhary, jens.bartelt, gerhard.fettweis}@tu-dresden.de

***Abstract*—Despite the promising benefits of the cloud-radio access network (C-RAN), the fronthaul (FH) imposes stringent requirements in terms of data rate, latency, jitter and synchronisation. In the classical C-RAN, the FH capacity scales linearly with the number of the transmitting antennas, which has posed severe demands on the FH capacity, especially due to emerging 5G technologies such as massive MIMO. However, this can be relaxed by performing precoding at the remote radio units (RRUs) instead of centrally, leading to FH traffic which depends on the number of currently served users. This paper adapts queueing theory results and spatial traffic models to derive statistical multiplexing gain enabled by varying number of user streams. Through this, we show that the required FH capacity can be reduced dramatically, depending on traffic demand and its statistical properties.**

***Index Terms*—Cloud-radio access network (C-RAN), fronthaul, backhaul, statistical multiplexing, outage probability**

I. Introduction and Motivation

In the recent years, demand of higher data and continuous connectivity along with the quality of experience (QoE) has grown exponentially and hence 5G promises to provide ubiquitous connection to everybody, anytime, anywhere. Moreover, future 5G wireless networks aims to support plethora of applications scenarios and use cases such as enhanced mobile broadband (eMBB) services, massive machine type communications (mMTC), and ultra-reliable and low latency communications (uRLLC) [1]. The traditional networks, where the BS performs the complete baseband processing including physical (PHY) layer, media access control (MAC) layer and a part of the network layer processing, suffer from several disadvantages such as higher capital expenditures (CAPEX) and operating expenses (OPEX). In order to cope with these challenges, cloud-radio access network (C-RAN) has been proposed [2], [3] that has drawn significant attention in the recent years. C-RAN enables efficient cooperative signal processing, and significantly lowers the network deployment and operation costs. In C-RAN, a part of the signal processing functionalities of traditional BSs are moved to the remote radio units (RRUs), which performs mostly the radio processing functionalities such as amplification, filtering etc. whereas the main digital processing is centralized to the baseband unit (BBU). The transport link between the RRU and BBU is known as FH and the FH typically uses common public radio interface (CPRI) [4]. The CPRI is a serial bidirectional digital interface that forwards the I/Q- (in-phase/quadrature-phase) samples from RRUs to central BBUs and vice versa. The FH commonly employs dark fiber, however, recently radio over fiber [5] or radio over Ethernet [6] are also employed.

Despite several advantages of C-RAN, one of the major challenges of C-RAN is to meet its high fronthaul capacity demand. Generally, the capacity of a fronthaul network is given by [7]

$$D_{FH} = 2 \cdot N_A \cdot f_s \cdot N_Q \cdot \gamma, \tag{1}$$

where N_A is number of antennas at the BS, f_s is sampling frequency, N_Q is quantization resolution bits of ADC or DAC, γ is overhead factor and the factor 2 accounts for I/Q samples. It is obvious from Eq. (1) that the FH capacity scales linearly with the number of transmitting antennas and hence with massive MIMO employing arrays of hundreds of transceivers, FH data rate increases several folds that an FH network with existing dedicated optical fiber can not support. For example, a 20 MHz LTE with 2×2 MIMO demands about 2.5 Gbps using CPRI line rate option 3 [4]. The recent CPRI specification even defines FH rates of up to 24 Gbps. Massive MIMO is a promising technology that employs hundreds of antenna at the BS and serves tens of users in the same time-frequency resources [8]–[10]. Massive MIMO offers several benefits such as high spectral efficiency (by simultaneously serving several users), high energy efficiency (by allowing multiple antennas to focus the radiated energy into even smaller regions of space), extensive use of inexpensive low-power components and significant reduction of latency on the air interface.

In order to meet the challenging requirement of FH data rate constraint, various approaches have been proposed such as (i) increasing FH capacity using wavelength division multiplexing (WDM) [11], time-shared optical networks (TSON) [12], [13], (ii) decreasing the FH capacity by using compressed signal processing [14], or functional splits between RRU-BBU [7], [15]. Furthermore, the FH capacity requirements could be decreased at the network level by allowing the several RRUs to multiplex data in the aggregation network [16], [17], [18, Chapter 4]. The direct link between the RRU and aggregation network is termed as *last mile* access, denoted by *FH Segment I* whereas the link between the aggregation network and the BBU is the main FH segment, denoted by *FH Segment II* as shown in Fig. 1. This figure shows that multiple RRUs are aggregated to the aggregation network and the resulting traffic

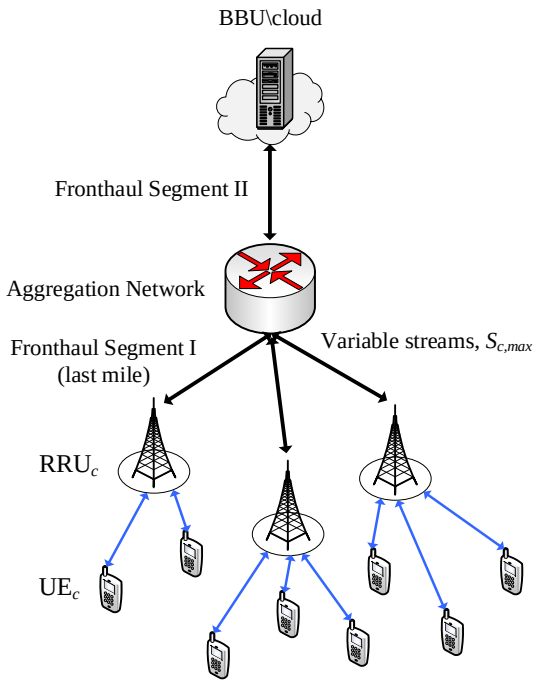


Fig. 1: Aggregation network showing multiplexing gain.

is then forwarded to the BBU via FH Segment II. Different topologies such as star, chain, ring and tree topologies can be used for FH as well as for backhaul (BH). The advantages of having FH Segment I will allow users to have data delivery with shorter cable lengths. On the other hand, FH Segment II demands to have higher capacity and more protection against the link failure.

The recent IEEE 1914 working group for next generation fronthaul interface (NGFI) [19] has adopted the network level solution, whereby with the help of switching functionalities multiple RRUs share a common FH link, thus enabling the multiplexing gain. Moreover, with the appropriate RRU-BBU functional split [7], [15], FH data rate can be made traffic dependent, unlike the traditional CPRI where FH data rate is always static and independent of the traffic load. Thus, the appropriate functional split will enable to transmit one stream per user instead of one stream per transceiver by performing MIMO precoding at the RRU instead of centrally [20]. This gives rise to two advantages: First, the number of streams will vary according to the users currently served. By allowing a certain outage probability within the limits of acceptable quality of service (QoS), i.e. dimensioning the FH capacity only for the 99$^{\text{th}}$ percentile of the traffic distribution, the required FH capacity can be reduced considerably. The second advantage of variable FH traffic is that in the aggregation segment, the streams of different RRUs will be combined, resulting in statistical multiplexing, which further lowers the required FH capacity. In this paper, we will describe a methodology how to evaluate these gains and quantify the benefits for different scenarios.

II. System Model

A. MIMO Rate

Let us consider downlink transmission of a massive MIMO system where the cellular network consists of a set $\mathcal{C}$ of C RRUs with index c, each serving a single cell. A RRU has M_c antennas that simultaneously serve K_c single-antenna users. Each cell has its own BS that can simultaneously serve K_{max} users, where K_{max} is the maximum number of users that a base station can serve at its peak load. Further, let us assume that the system operates in time division duplexing (TDD) mode such that channel is reciprocal and the BS estimates the downlink channels by using uplink received pilots [8]. The number of pilots is limited to K_{max}, which also means that each RRU can serve at maximum K_{max} users simultaneously.

The area served by the RRUs is denoted $\mathcal{A}$, with a single location indicated by its coordinates (x, y). The pathloss between RRU c and location (x, y) is denoted $\alpha_c(x, y)$ and modelled according to the urban microcellular pathloss defined in [21].

User are associated with the RRU providing the lowest pathloss, hence the serving area $\mathcal{A}_c$ of a RRU c is given as

$$(x, y) \in \mathcal{A}_c \text{ if } c = \arg\max_{\mathcal{C}} \alpha_c(x, y). \tag{2}$$

Let us consider that the total transmit power of a BS is pM_c, where is p is the average power per antenna that is considered to be same for all the antennas. Each cell c receive interference from the active antennas in the any other cell d. Let M_d be the active antennas in any other cell, hence total transmitting power of the corresponding cell is pM_d. Then, the SINR at a location is given as:

$$\gamma(x, y) = \frac{pM_c\alpha_b(x, y)}{\sigma^2 + \sum_{d\in\mathcal{C}\setminus c} pM_d\alpha_i(x, y)}, \tag{3}$$

where σ^2 denotes the noise power.

We use the notion of Poisson point processes to describe the traffic demand in the network, compare e.g. [22]. For this, we define for each location a user arrival rate per area $\Lambda(x, y)$ (in $1/\text{s}/\text{km}^2$) and hence a corresponding traffic density (in Mbps/km^2) is

$$\Omega(x, y) = \Lambda(x, y) \cdot s, \tag{4}$$

where s (in Megabit) is the mean file size requested per user. The mean traffic density of the overall area $\mathcal{A}$ we denote $\bar{\Omega}$. For the serving area of a RRU this results in user arrivals with arrival rate (in $1/\text{s}$)

$$\lambda_c = \int_{\mathcal{A}_c} \Lambda(x, y)\, dx\, dy. \tag{5}$$

From this, we define the average SINR in the serving area of RRU c as the expected value of the SINRs weighted according to the traffic distribution $\Omega(x, y)$, i.e.:

$$\bar{\gamma}_c = \mathbb{E}[\gamma(x, y)] = \frac{\int_{\mathcal{A}_c} \gamma_c(x, y)\Omega(x, y)\, dx\, dy}{\int_{\mathcal{A}_c} \Omega(x, y)\, dx\, dy}. \tag{6}$$

Furthermore, for simplicity let us assume that each RRU has

obtained perfect CSI from its users and employs zero forcing precoding in order to cancel out the intracell interference and adapts power allocation such that each of the K_c user achieves the same average data rate, R_c given by [23]:

$$R_c(K_c) = B\left(1 - \frac{K_{max}}{T_c}\right)\log_2\left(1 + \frac{\bar{\gamma}}{K_c}(M_c - K_c)\right), \tag{7}$$

where B is channel bandwidth, $T_c = B_{coh}\tau_{coh}$ is length of channel coherence interval, B_{coh} is coherence bandwidth, τ_{coh} is coherence time, and K_{max} is the maximum number of users, which is assumed to be same for all cells. The pre-log factor $\left(1 - \frac{K_{max}}{T_c}\right)$ is channel estimation overhead, $M_c - K_c$ is the effective array gain, and the factor $1/K_c$ accounts for the fact that the total transmit power is split between all users. Furthermore, the above formula has constraints $M_c - K_c \geq 1$ and $d \neq c$ that arises due to requirement of zero forcing precoding [23].

An example of this rate is given in Fig. 2, which shows the average data rate per user R_c, and Fig. 3, which shows the total sum throughput $K_c \cdot R_c(K_c)$. It is obvious from Fig. 2 that higher average rate per user is achievable by having fewer active users. On the other hand, the Fig. 3 shows that the sum throughput increases when increasing the number of active users. The simulation parameters for all the figures are listed in Table I.

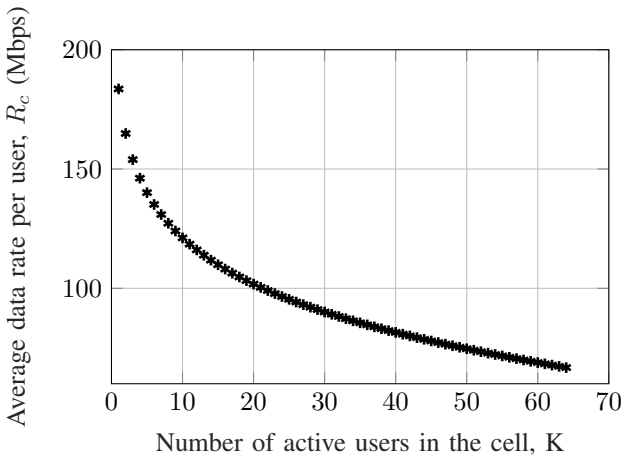


Fig. 2: Average data rate per user.

B. Queuing model

In order to evaluate outage probabilities and statistical multiplexing, we need to find the number of active users in a cell. For this, we utilize queueing theory results from [24], [25], where each MIMO RRU is modeled as $M/G/m/m$ state-dependent queue. The $M/G/m/m$ queue states that for exponential arrival and general distribution of service time, maximum m number of users can be served simultaneously and the arrivals follow the Poisson process [23]. In this work, we assume that the maximum number of users that a BS can serve is $m = K_{max}$. Let $\pi_c(n) \equiv P_r[K_c = n]$ be the steady

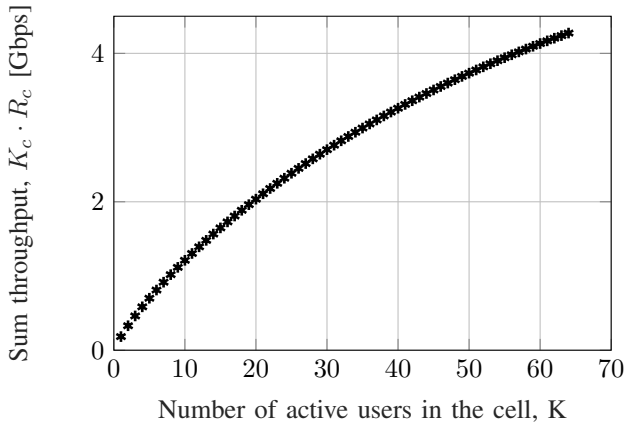


Fig. 3: Sum throughput of the users.

state probabilities of having K_c active users served by RRU c, then under $M/G/m/m$ state-dependent queue, $\pi_c(n)$ is given by [24]:

$$\pi_c(n) = \left[\frac{\left[\lambda_c \frac{s}{R_c(1)}\right]^n}{n!f(n)f(n-1)\dots f(2)f(1)}\right]\pi_c(0), n = 1, 2, \dots m, \tag{8}$$

with

$$\pi_c^{-1}(0) = 1 + \sum_{i=1}^{m}\left(\frac{\left[\lambda_c \frac{s}{R_c(1)}\right]^i}{i!f(i)f(i-1)\dots f(2)f(1)}\right),$$

where $f(n) = R_c(n)/R_c(1)$ is the normalized rate per user, R_c is the average data rate per user while serving n number of users given by Eq. (7), λ_c is the arrival rate from Eq. (5), and s is the average file size requested by each user introduced earlier. The corresponding CDF is denoted as $\Pi_c(n) = \sum_{i \leq n}\pi_c(i)$. An example of these user distributions is given in Fig. 4 for different values of λ_c. This figure implies that the as the arrival rate increases, there is more flows per second from the users and the number of users attempting to get the resources is also increasing. For example, at 2% probability, the number of active users for arrival rate of $\lambda_c = 5$ flows per second is 12 which increases to 28 users for arrival rate of $\lambda_c = 25$ flows per second.

C. Outage and Multiplexing

Conventionally, the FH in Segment I would be dimensioned to serve its maximum number of users, i.e. for a number of stream $S_{c,\text{max}} = K_{\text{max}}$ streams. Similarly, Segment II would be dimensioned for $S_{\mathcal{C}} = C \cdot K_{\text{max}}$ streams. Such a dimensioning is common in current CPRI-based FH networks, which required a constant data rate. This constant data rate in the FH is what makes the FH-constrained, which eventually will be bottleneck for massive MIMO systems. However, due to the variability in traffic due to the varying number of user-streams, we can assume a certain outage probability P_{O} on

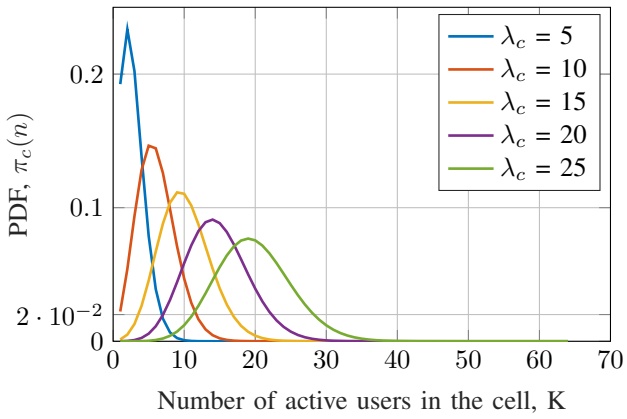


Fig. 4: Arrival rates.

each link according to some QoS requirements. Hence, the FH in Segment I can be dimensioned with the outage capacity

$$S_{c,\mathrm{O}} = n \text{ such that } \Pi_c(n) < P_{\mathrm{O}}. \tag{9}$$

Furthermore, the streams are aggregated for Segment II, i.e. the number of streams are summed up. The summation leads to a convolution of the corresponding probability distributions, i.e.

$$\pi_{\mathcal{C}} = \pi_1 * \pi_2 * \cdots * \pi_c, \tag{10}$$

with CDF $\Pi_{\mathcal{C}}$ and outage capacity

$$S_{\mathcal{C},\mathrm{O}} = n \text{ such that } \Pi_{\mathcal{C}}(n) < P_{\mathrm{O}}. \tag{11}$$

An example of this is illustrated in Figs. 5a and 5. The total traffic from 19 such cells, assuming each cell in it peak load can serve 64 users, demands to have total $64 \times 19 = 1216$ user data to be forwarded. However, assuming a reasonable 1% outage probability of the fronthaul segment II, we need to transport only 605 users as shown in Fig. 5, which means less fronthaul capacity demand. It allows up to 50% FH capacity saving.

In general, the convolution will lead to a a longer-tailed PDF, which yields to a statistical mutiplexing gain, as $S_{\mathcal{C},\mathrm{O}} \leq \sum_{\mathcal{C}} S_{c,\mathrm{O}}$. To assess the benefit of the statistical multiplexing, we define the relative required FH rates in Segment I and 2 as:

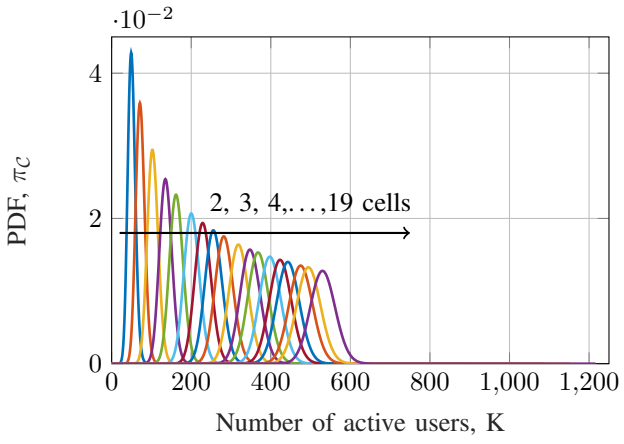
$$S_1 = \frac{\sum_{\mathcal{C}} S_{c,\mathrm{O}}}{C \cdot S_{c,\max}} \qquad \text{for FH Segment I,} \tag{12}$$

$$S_2 = \frac{S_{\mathcal{C},\mathrm{O}}}{C \cdot S_{c,\max}} \qquad \text{for FH Segment II.} \tag{13}$$

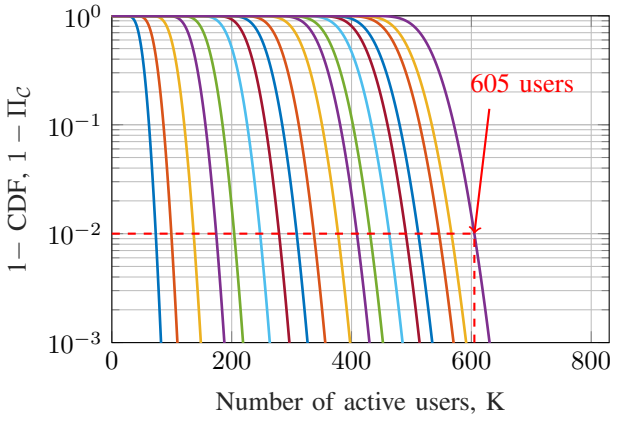
In summary, we can observe two methods to lower the required FH capacity: first, accepting a certain outage (which benefits both FH Segments 1 and 2), and second, accounting for the effect of statistical multiplexing in the aggregation part (which is only possible in Segment II).

D. Traffic Model

In general, the gain of statistical multiplexing will depend on the variance of the total number of streams. This variance



(a) PDF.



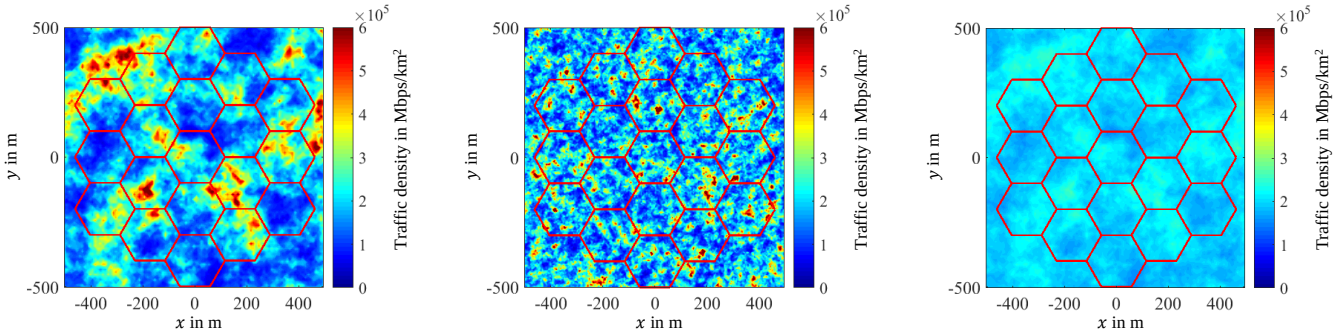
(b) CCDF.

Fig. 5: Exemplary distribution of served users/required user streams for 19 aggregated RRUs. $\bar{\Omega} = 45$ Gbps/km^2, $\sigma_\Omega = 0.25\ \bar{\Omega}$, $d_{\mathrm{corr}} = 10$ m.

is affected both by the (temporal) variation fo users from Eq. (8), and by the different (spatial) variation of users among different cells based on the traffic distribution $\Omega(x, y)$. In order to model $\Omega(x, y)$, we utilize a traffic model developed in [26], [27]. This traffic model allows to create random spatial traffic maps via log-normal distributed random fields defined by three statistical parameters. These parameters are the mean traffic density, $\bar{\Omega}$, the traffic density standard deviation, σ_Ω, and a correlation density d_{corr}. Three examples of such traffic maps are given in Fig. 6. The parameter $\bar{\Omega}$ controls the overall traffic demand, σ_Ω controls the ration between traffic demand in hot spots and low-traffic areas, and d_{corr} controls the size of the hotspots. With traffic maps generated based on this model, we can average statistical multiplexing gains over random scenarios without having to rely on just a single scenario, leading to more consistent results and more general conclusions for real scenarios. For more details on this traffic model see [27].

III. Numerical Results

To evaluate the FH capacity reduction, we utilize an exemplary setup consisting of 19 uniformly placed hexagonal



(a) $\bar{\Omega} = 200$ Gbps/km^2, $\sigma_\Omega = 100$ Gbps/km^2, $d_{\text{corr}} = 50$ m.

(b) $\bar{\Omega} = 200$ Gbps/km^2, $\sigma_\Omega = 100$ Gbps/km^2, $d_{\text{corr}} = 10$ m.

(c) $\bar{\Omega} = 200$ Gbps/km^2, $\sigma_\Omega = 20$ Gbps/km^2, $d_{\text{corr}} = 50$ m.

Fig. 6: Exemplary traffic distributions. Fig. (b) exhibits a lower correlation distance, and Fig. (c) a lower standard deviation compared to Fig. (a).

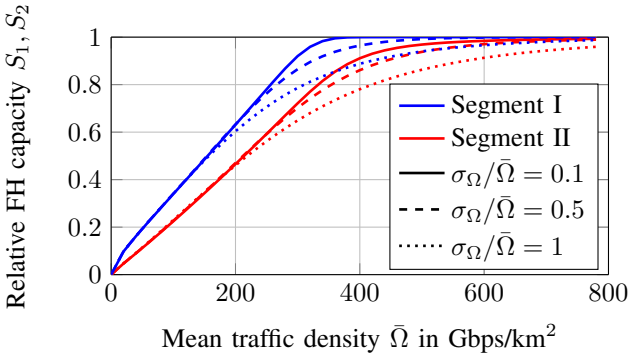


Fig. 7: Relative FH capacity for different relative standard deviations of the traffic density, $d_{\text{corr}} = 50$ m, $P_{\text{O}} = 0.01$.

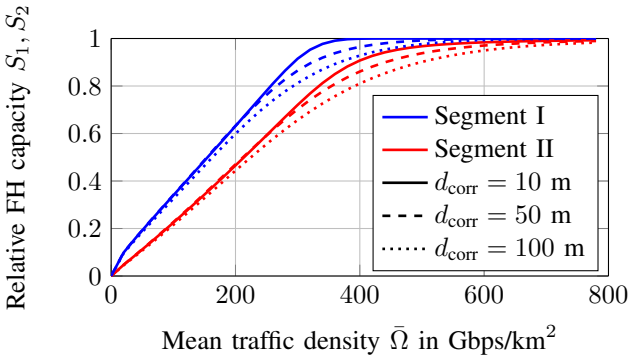


Fig. 8: Relative FH capacity for different traffic correlation distances, $\sigma_\Omega/\Omega = 0.5$, $P_{\text{O}} = 0.01$.

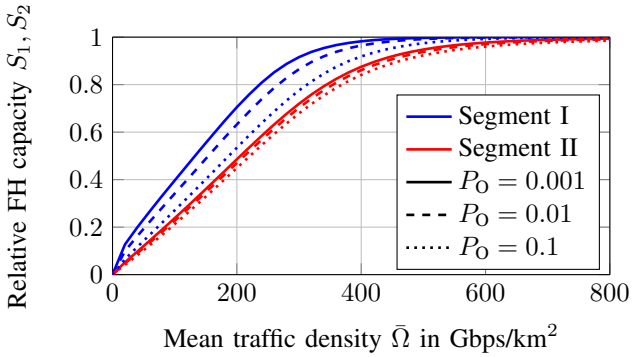


Fig. 9: Relative FH capacity for different outage probabilities, $\sigma_\Omega/\Omega = 0.5$, $d_{\text{corr}} = 50$ m.

TABLE I: Simulation parameters.

Parameters	**Symbol**	**Value**
Number of cells	C	19
Inter site distance	d_{ISD}	200 m
RRH height	h_{RRU}	12 m
Bandwidth	B	20 MHz
Coherence bandwidth	B_{coh}	200 kHz
Coherence time	T_{coh}	5 ms
Number of transmitting antennas	M_c	256
Maximum users	K_{max}	64
Total transmit power	pM_c	23 dBm
Total noise power	σ^2	-96 dBm
Average file size	s	80 Mbit

cells with inter site distance $d_{\text{ISD}} = 200$ m and RRUs placed at a height of $h_{\text{RRU}} = 12$ m. These are placed on random traffic maps generated according to Sec. II-D, which is also illustrated in Figs. 7-9. Then, the relative required FH rates according to Eqs. (12), (13) are evaluated and the results are averaged over 25 random traffic maps.

The results are illustrated in Figs. 7-9. As can be seen, the relativ FH capacity mainly scales with the mean traffic density. In addition, the capacity in FH Segment II is always lower, as here the additional effect of statistical multiplexing comes into effect. The difference between Segment I and 2 is more pronounced towards higher traffic densities. Here clearly the statistical multiplexing effect is more dominant compared to the reduction possible by accepting outage. Furthermore, it can be seen that higher values of traffic variance and correlation distance lead to lower FH capacities, as both parameters lead

to a higher variability in total cell traffic among the different RUs, hence resulting in a higher multiplexing gain. Finally, it can be seen in Fig. 9 that a higher outage probability leads to an lower required FH capacity, as can be expected. Here, especially Segment I profits. In Segment II, due to the statistical multiplexing effect, the probability distribution converges towards the mean traffic, and hence the difference between the percentiles are less pronounced.

IV. Conclusion

In spite of several benefits of C-RAN, massive MIMO systems have posed severe challenges on its fronthaul (FH) capacity leading to FH being constrained. In this paper, we have adapted queuing theory and traffic model to analyse network level solution via statistical multiplexing to relax FH capacity. Further, we think the method we have adopted will provide newer insights to analyse statistical multiplexing using queuing theory and traffic model. We showed that considerable amount of FH capacity can be saved by allowing multiplexing gain with certain outage probability within the limits of acceptable quality of service.

Acknowledgment

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 671551 (5G-XHaul). The European Union and its agencies are not liable or otherwise responsible for the contents of this document; its content reflects the views of its authors only.

References

[1] NGMN Alliance, "Ngmn white paper," March 2015, online: http://www.ngmn.org.

[2] China Mobile Research Institute, "C-RAN - The road towards green RAN," *White Paper*, Oct. 2011, [Online]. Available: http://labs.chinamobile.com/cran/wp-content/uploads/CRAN_white_paper_v2_5_EN.pdf. Accessed Sep. 2, 2016.

[3] A. Checko, H. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. Berger, and L. Dittmann, "Cloud ran for mobile networks; a technology overview," *Communications Surveys Tutorials, IEEE*, vol. 17, no. 1, pp. 405–426, Firstquarter 2015.

[4] CPRI, "Common Public Radio Interface (CPRI); Interface Specification (V7.0)," Tech. Rep., Oct. 2015, online: http://www.cpri.info/.

[5] M. Sauer, A. Kobyakov, and J. George, "Radio over fiber for picocellular network architectures," *Journal of Lightwave Technology*, vol. 25, no. 11, pp. 3301–3320, Nov 2007.

[6] IEEE 1904.3 Task Force, "Standard for radio over ethernet encapsulations and mappings," online: http://grouper.ieee.org/groups/1904/.

[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, June 2013.

[8] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Transactions on Wireless Communications*, vol. 9, no. 11, pp. 3590–3600, November 2010.

[9] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up mimo: Opportunities and challenges with very large arrays," *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 40–60, Jan 2013.

[10] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive mimo for next generation wireless systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 186–195, February 2014.

[11] 5G-XHaul Project, Deliverable D2.1, "Requirements specification and kpis document," [Online]. Available: http://http://www.5g-xhaul-project.eu/download/5G-XHaul_D_21.pdf. Accessed Feb. 10, 2017.

[12] G. S. Zervas, J. Triay, N. Amaya, Y. Qin, C. Cervell-Pastor, and D. Simeonidou, "Time shared optical network (tson): A novel metro architecture for flexible multi-granular services," in *2011 37th European Conference and Exhibition on Optical Communication*, Sept 2011, pp. 1–3.

[13] Y. Yan, Y. Qin, G. Zervas, B. Rofoee, and D. Simeonidou, "High performance and flexible fpga-based time shared optical network (tson) metro node," in *2012 38th European Conference and Exhibition on Optical Communications*, Sept 2012, pp. 1–3.

[14] J. Lorca and L. Cucala, "Lossless compression technique for the fronthaul of lte/lte-advanced cloud-ran architectures," in *2013 IEEE 14th International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)*, June 2013, pp. 1–9.

[15] D. Wübben, P. Rost, J. S. Bartelt, M. Lalam, V. Savin, M. Gorgoglione, A. Dekorsy, and G. Fettweis, "Benefits and impact of cloud computing on 5g signal processing: Flexible centralization through cloud-ran," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 35–44, Nov 2014.

[16] J. Liu, S. Xu, S. Zhou, and Z. Niu, "Redesigning fronthaul for next-generation networks: beyond baseband samples and point-to-point links," *IEEE Wireless Communications*, vol. 22, no. 5, pp. 90–97, October 2015.

[17] J. Bartelt, P. Rost, D. Wubben, J. Lessmann, B. Melis, and G. Fettweis, "Fronthaul and backhaul requirements of flexibly centralized radio access networks," *IEEE Wireless Communications*, vol. 22, no. 5, pp. 105–111, October 2015.

[18] K. M. S. Huq and J. Rodriguez, *Backhauling/Fronthauling for Future Wireless Systems*. John Wiley & Sons, 2016.

[19] "Next Generation Fronthaul Interface," http://sites.ieee.org/sagroups-1914/.

[20] S. Park, C. B. Chae, and S. Bahk, "Before/after precoding massive mimo systems for cloud radio access networks," *Journal of Communications and Networks*, vol. 15, no. 4, pp. 398–406, Aug 2013.

[21] ITU-R, *ITU Recommendation ITU-R M.2135-1: Guidelines for evaluation of radio interface technologies for IMT-Advanced*, ITU-R Std., Dec. 2009.

[22] F. Baccelli and B. Blaszczyszyn, *Stochastic Geometry and Wireless Networks, Volume I - Theory*, ser. Foundations and Trends in Networking Vol. 3: No 3-4, pp 249-449, F. Baccelli and B. Blaszczyszyn, Eds. NoW Publishers, 2009, vol. 1, stochastic Geometry and Wireless Networks, Volume II - Applications; see http://hal.inria.fr/inria-00403040. [Online]. Available: https://hal.inria.fr/inria-00403039

[23] M. M. A. Hossain, C. Cavdar, E. Bjornson, and R. Jantti, "Energy-efficient load-adaptive massive mimo," in *2015 IEEE Globecom Workshops (GC Wkshps)*, Dec 2015, pp. 1–6.

[24] F. Cruz and J. M. Smith, "Approximate analysis of m/g/c/c state-dependent queueing networks," *Computers and Operations Research*, vol. 34, no. 8, pp. 2332 – 2344, 2007. [Online]. Available: //www.sciencedirect.com/science/article/pii/S0305054805002960

[25] J. Y. Cheah and J. M. Smith, "Generalized m/g/c/c state dependent queueing models and pedestrian traffic flows," *Queueing Systems*, vol. 15, no. 1, pp. 365–386, 1994. [Online]. Available: http://dx.doi.org/10.1007/BF01189246

[26] D. Lee, S. Zhou, X. Zhong, Z. Niu, X. Zhou, and H. Zhang, "Spatial modeling of the traffic density in cellular networks," *IEEE Wireless Communications*, vol. 21, no. 1, pp. 80–88, February 2014.

[27] H. Klessig, M. Soszka, and G. Fettweis, "Multi-cell flow-level performance of traffic-adaptive beamforming under realistic spatial traffic conditions," in *2015 International Symposium on Wireless Communication Systems (ISWCS)*, Aug 2015, pp. 726–730.