Federation University ResearchOnline

This is the peer-reviewed version of the following article:

Vu, T. T., Ngo, H. Q., Dao, M. N., Matthaiou, M., & Larsson, E. G. (2022). Data Size-Aware Downlink Massive MIMO: A Session-Based Approach. *IEEE Wireless Communications Letters, 11*(7), 1–1.

Available online: https://doi.org/10.1109/LWC.2022.3174829

See this record in Federation ResearchOnline at:
http://researchonline.federation.edu.au/vital/access/HandleResolver/1959.17/191898

Data Size-Aware Downlink Massive MIMO: A Session-Based Approach

Tung T. Vu, Hien Quoc Ngo, Minh N. Dao, Michail Matthaiou, and Erik G. Larsson

Abstract—This letter considers the development of transmission strategies for the downlink of massive multiple-input multiple-output networks, with the objective of minimizing the completion time of the transmission. Specifically, we introduce a session-based scheme that splits time into sessions and allocates different rates in different sessions for the different users. In each session, one user is selected to complete its transmission and will not join subsequent sessions, which results in successively lower levels of interference when moving from one session to the next. An algorithm is developed to assign users and allocate transmit power that minimizes the completion time. Numerical results show that our proposed session-based scheme significantly outperforms conventional non-session-based schemes.

Index Terms—Massive MIMO, session-based, zero-forcing.

I. INTRODUCTION

We are witnessing an explosion of streaming and learning applications, such as video streaming, live conferencing, and federated learning [1]–[3]. Many of these applications require computations by mobile users (UEs) [4], and the UEs have fixed amounts of data to receive. To support these applications, it is critical to design transmission schemes that achieve low latency. It is of particular importance to design communication protocols that minimize the completion time, defined as the time it takes for a UE to receive all data destined for it.

To support the aforementioned applications, massive multiple-input multiple-output (MIMO) can be used due to its ability to offer high data rates to all UEs simultaneously [5]. To reduce the completion time, conventionally, the achievable rates of all the UEs are maximized via power allocation, and these rates are kept constant during the whole transmission. Another approach is to use different rates during the transmission period [6]. When some UEs have already completed their transmissions, other UEs will benefit by having less multi-user interference and higher rates, which results in shorter completion times. In this context, [6] studies the completion time for two-user systems from an information-theoretic perspective. However, general schemes for multi-user systems that use different rates within the transmission period have not been explored in the literature.

Contributions: Motivated by [6], we introduce a session-based scheme to reduce the completion times of UEs for the downlink of massive MIMO networks. In this scheme, UE

T. T. Vu, H. Q. Ngo, and M. Matthaiou are with the Centre for Wireless Innovation (CWI), Queen's University Belfast, Belfast BT3 9DT, UK (e-mail: {t.vu, hien.ngo,m.matthaiou}@qub.ac.uk).

M. N. Dao is with the School of Engineering, Information Technology and Physical Sciences, Federation University, VIC 3353, Australia (e-mail: m.dao@federation.edu.au).

Erik G. Larsson is with Department of Electrical Engineering (ISY), Linköping University, SE-581 83 Linköping, Sweden (e-mail: erik.g.larsson@liu.se).

data is transmitted with different rates in different sessions. Specifically, in each session, UEs are assigned so that one UE finishes its transmission and does not participate in subsequent sessions. UEs with uncompleted transmissions are allocated more power to obtain higher achievable rates, and complete their transmissions faster. Herein, zero-forcing (ZF) processing is used for data transmission. An algorithm is developed for assigning UEs and allocating transmit power, with the objective of minimizing the completion times of the UEs. Numerical results show that the proposed session-based scheme significantly reduces the completion times compared to conventional transmission that relies on power control only.

A specific version of this session-based scheme was also used in [7], although for a different objective and for a particular application. Herein, we substantially extend [7] to the general problem of minimizing downlink completion times. If the amounts of UE data in each session are given, then the optimization problem reduces to power and rate control for conventional transmission (with different rate constraints for different users). However, optimizing the amounts of data per session and the thresholds for the rate constraints over multiple sessions is a new and challenging problem.

II. SYSTEM MODEL

We consider the downlink transmission in a massive MIMO network, where an M-antenna base station (BS) serves $K \leq M$ single-antenna UEs simultaneously in the same frequency band. Let S_k be the size of the data intended for UE k. We focus on applications where S_k is fixed, such as mobile edge computing and federated learning to name but a few [3], [8]. We assume that the transmission time is within one large-scale coherence time,[1] and the transmission is spans multiple small-scale coherence blocks. A small-scale coherence block is the time-frequency interval over which the channel is substantially static, and is divided into two phases: channel estimation and downlink payload data transmission.

Resource allocation, such as transmit power control, is typically performed to guarantee given quality-of-service targets. We consider two conventional schemes: (i) **Conventional non-data size-aware scheme**: This scheme allocates power such that all users achieve the same rate [5]; and (ii) **Conventional data size-aware scheme**: This scheme allocates power such that each UE receives a rate proportional to the size of its data. This is the traditional scheme used in wireless networks supporting mobile edge computing or federated learning (see, e.g., [3], [8] and references therein).

In both conventional schemes, the data rates are kept fixed for the whole transmission. However, since K UEs have different required data sizes, some UEs could complete their transmissions before other UEs. Therefore, optimally, the data rates vary temporally depending on how many active UEs remain in the

[1]The large-scale coherence time is the time during which the large-scale fading coefficients remain substantially constant [3].

system.[2] The main question is how to allocate power and update the UEs' rates to reduce the completion times. Motivated by this, we next propose the novel session based-scheme.

III. PROPOSED SESSION-BASED SCHEME

Our proposed scheme uses different rates in different time periods for data transmissions. The transmission period during which the rates are kept fixed is called a "session". More precisely, session i is defined as follows: (i) in session i, there are $K-i+1$ active UEs; and (ii) at the beginning of session i, the BS updates the rates for these active UEs. These rates will be kept fixed until the end of this session, where one UE completes receiving data from the BS. Thus, the BS transmits data to all K UEs during K sessions. Fig. 1 illustrates the proposed session-based transmission as well as conventional transmission for a system with three UEs. Denote by a_i the indicator that is defined as

$$a_{k,i} \triangleq \begin{cases} 1, & \text{if UE } k \text{ is receiving data in session } i, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

Denote by $\mathcal{K}_i \triangleq \{k|a_{k,i}=1\}$ the set of $K_i = \sum_{k \in \mathcal{K}} a_{k,i}$ UEs assigned in session $i \in \mathcal{K}$, where $\mathcal{K} = \{1, 2, \ldots, K\}$. Then, we have

$$a_{k,1}=1, \sum_{k \in \mathcal{K}} a_{k,i} = K-i+1, a_{k,i} \le a_{k,i-1}, \forall i, \quad (2)$$

to ensure that all the UEs will be served in session 1. In each subsequent session, one UE completes its transmission and will not join the next sessions. As such, more power is allocated to the UEs that have not yet completed their transmissions. Note that, the conventional (non-session-based) schemes are special cases of the proposed scheme when all UEs are served in session 1 with $a_{k,1}=1, \forall k$.

Uplink channel estimation: In each small-scale coherence block of length τ_c, each UE sends its pilot of length τ_p to the BS. We assume that all pilots are mutually orthogonal, which requires $\tau_p \ge K$.[3] Denote by $\boldsymbol{g}_k = (\beta_k)^{1/2} \tilde{\boldsymbol{g}}_k$ the channel vector from UE k to the BS, where β_k and $\tilde{\boldsymbol{g}}_k$ are the large-scale fading coefficient and small-scale fading vector, respectively. With minimum mean-square error (MMSE) estimation, the channel estimate $\hat{\boldsymbol{g}}_k$ of $\boldsymbol{g}_k$ is distributed according to $\mathcal{CN}(\boldsymbol{0}, \sigma_k^2 \boldsymbol{I}_M)$, where $\sigma_k^2 = \frac{\tau_p \rho_p \beta_k^2}{\tau_p \rho_p \beta_k + 1}$, and ρ_p is the normalized transmit power of each pilot symbol [5, (3.8)]. Let $\hat{\boldsymbol{G}}_i \triangleq [\hat{\boldsymbol{g}}_1, \ldots, \hat{\boldsymbol{g}}_{K-i+1}], \forall k \in \mathcal{K}_i$, be the matrix stacking the estimated channels of all participated UEs in session i.

Downlink payload data transmission: In session i, the BS uses ZF to transmit data to $K-i+1$ UEs. With ZF, the signal transmitted by the BS is given by $\boldsymbol{x}_i = \sqrt{\rho} \sum_{k \in \mathcal{K}_i} \boldsymbol{u}_{k,i} s_{k,i}$, where ρ is the normalized transmit power at the BS; $s_{k,i}$, with $\mathbb{E}\{|s_{k,i}|^2\} = 1$, is the symbol intended for UE k; and $\boldsymbol{u}_{k,i} = \sqrt{\eta_{k,i} \sigma_k^2 (M-K_i)} \hat{\boldsymbol{G}}_i (\hat{\boldsymbol{G}}_i^H \hat{\boldsymbol{G}}_i)^{-1} \boldsymbol{e}_{k,K_i}$ is the ZF precoding vector. Here, $\mathbb{E}\{x\}$ denotes the expected value of a random variable x, and $\boldsymbol{X}^H$ represents the conjugate transpose of a matrix $\boldsymbol{X}$. In the precoding vector, $\eta_{k,i}$ is a power control coefficient, and $\boldsymbol{e}_{k,K_i}$ is the k-th column of $\boldsymbol{I}_{K_i}$. The transmitted

---

[2]A UE that completely receives data from the BS will become inactive.

[3]One can let only the participating UEs send their pilots in session i, i.e., $\tau_p = K_i$ to increase the small-scale coherence block length for payload data transmission. However, since τ_c is normally much larger than $K \ge K_i$ in many applications [3], letting all the UEs send their pilots, i.e., taking $\tau_p = K$, has a negligible impact on data rates. On the other hand, the channel estimation is better when the pilot length $\tau_p = K > K_i$, which potentially improves the data rates.
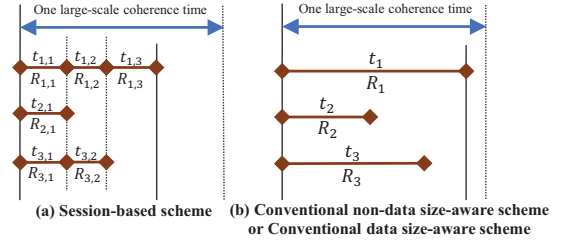

(a) Session-based scheme   (b) Conventional non-data size-aware scheme or Conventional data size-aware scheme

Fig. 1. One transmission period with three UEs.

power at the BS is constrained by $\mathbb{E}\{|\boldsymbol{x}_i|^2\} \le \rho$ which is equivalent to

$$\sum_{k \in \mathcal{K}_i} \eta_{k,i} \le 1, \forall i \in \mathcal{K}. \quad (3)$$

We enforce

$$(\eta_{k,i}=0, \text{if } a_{k,i}=0), \forall k,i \quad (4)$$

to ensure that the BS will not allocate any power to UEs that are not served in session i. The achievable rate at UE k in session i is given by [5, Eq. (3.56)]: $R_{k,i}(\boldsymbol{\eta}_i) = \frac{\tau_c - \tau_p}{\tau_c} B \log_2 \left(1 + \text{SINR}_{k,i}(\boldsymbol{\eta}_i)\right)$, where B is the bandwidth, $\text{SINR}_{k,i}(\boldsymbol{\eta}_i) = \frac{(M-K_i)\rho \hat{\sigma}_k^2 \eta_{k,i}}{\rho(\beta_k - \sigma_k^2) \sum_{\ell \in \mathcal{K}_i} \eta_{\ell,i} + 1} \stackrel{(4)}{=} \frac{(M-K_i)\rho \sigma_k^2 \eta_{k,i}}{\rho(\beta_k - \hat{\sigma}_k^2) \sum_{\ell \in \mathcal{K}} \eta_{\ell,i} + 1}$ is the effective downlink signal-to-interference-plus-noise ratio (SINR), and $\boldsymbol{\eta}_i \triangleq \{\eta_{k,i}\}_{k \in \mathcal{K}}$.

Completion time: Let $S_{k,i}$ be the size of the data sent to UE k in session i. Then, we have

$$\sum_{i \in \mathcal{K}} S_{k,i} = S_k, \forall k. \quad (5)$$

Let t_i be the time duration of session i. Then, the transmission time $t_{k,i}$ of UE $k \in \mathcal{K}$ in session i is given by

$$t_{k,i} = a_{k,i} t_i, \forall k,i. \quad (6)$$

Thus,

$$S_{k,i} = R_{k,i}(\boldsymbol{\eta}_i) t_{k,i} \stackrel{(6)}{=} R_{k,i}(\boldsymbol{\eta}_i) a_{k,i} t_i \stackrel{(4)}{=} R_{k,i}(\boldsymbol{\eta}_i) t_i, \forall k,i. \quad (7)$$

Clearly, (7) also implies that $(S_{d,k,i}=0, \text{if } a_{k,i}=0), \forall k,i$. The completion time of UE k is the sum of its transmission times across all sessions, i.e., $\sum_{i \in \mathcal{K}} a_{k,i} t_i$. Let $\tilde{T}_c$ and T_c the large-scale and small-scale coherence times, respectively. Since each session spans multiple small-scale coherence blocks but always fits within one large-scale coherence time, we have

$$T_c \le t_i, \forall i \quad (8)$$

$$\sum_{i \in \mathcal{K}} t_i \le \tilde{T}_c. \quad (9)$$

Remark 1. *In this work, in order to focus on fundamental principles of the proposed scheme, that is, UE assignment and rate allocation, we consider independent Rayleigh channels. The optimization and analysis of a session-based scheme tailored to correlated channels is interesting, but analytically challenging and beyond the scope of the paper. Thus, such designs are left for future work.*

IV. COMPLETION TIME MINIMIZATION

We address the minimization of the completion time of our proposed session-based scheme, and specifically minimizing the longest completion time among the UEs. For comparison, we also include the completion time minimization for the conventional schemes.

A. Proposed Session-Based Scheme

The problem of minimizing the completion time of UEs by optimizing the UE assignment ($\boldsymbol{a}$) and transmit power ($\boldsymbol{\eta}$) in the session-based design is

$$\min_{\boldsymbol{x}} \max_k \sum_{i \in \mathcal{K}} a_{k,i} t_i, \quad (10a)$$

s.t. $(1)-(4),(5),(7)-(9)$

$$0 \le \eta_{k,i}, \forall k,i, \tag{10b}$$

where $\boldsymbol{x} \triangleq \{\boldsymbol{a},\boldsymbol{\eta},\boldsymbol{S},\boldsymbol{t}\}$, $\boldsymbol{S} \triangleq \{S_{k,i}\}$, $\boldsymbol{a} \triangleq \{a_{k,i}\}$, $\boldsymbol{t} \triangleq \{t_i\}$ $\forall k,i$.

Finding a globally optimal solution to problem (10) is challenging due to the mixed-integer and nonconvex constraints (1), (4), and (7). Thus, we instead propose an approach that is suitable for practical implementation. First, we replace constraint (4) by

$$\eta_{k,i} \le a_{k,i}, \forall k,i, \tag{11}$$

and constraint (7) by

$$S_{k,i} \le R_{k,i}(\boldsymbol{\eta}_i)a_{k,i}t_i, \forall k,i \tag{12}$$

$$S_{k,i} \ge R_{k,i}(\boldsymbol{\eta}_i)a_{k,i}t_i, \forall k,i. \tag{13}$$

We further replace constraints (12) and (13) by

$$\hat{r}_{k,i} \le R_{k,i}(\boldsymbol{\eta}_i), \forall k,i \tag{14}$$

$$\tilde{r}_{k,i} \ge R_{k,i}(\boldsymbol{\eta}_i), \forall k,i \tag{15}$$

$$\hat{t}_{k,i} \le a_{k,i}t_i, \forall k,i \tag{16}$$

$$\tilde{t}_{k,i} \ge a_{k,i}t_i, \forall k,i \tag{17}$$

$$S_{k,i} \le \hat{r}_{k,i}\hat{t}_{k,i}, \forall k,i \tag{18}$$

$$S_{k,i} \ge \tilde{r}_{k,i}\tilde{t}_{k,i}, \forall k,i, \tag{19}$$

where $\hat{\boldsymbol{r}} \triangleq \{\hat{r}_{k,i}\}, \tilde{\boldsymbol{r}} \triangleq \{\tilde{r}_{k,i}\}, \hat{\boldsymbol{t}} \triangleq \{\hat{t}_{k,i}\}, \tilde{\boldsymbol{t}} \triangleq \{\tilde{t}_{k,i}\}$ are additional variables. We observe from (14)–(18) that $S_{k,i} \le \tilde{r}_{k,i}\tilde{t}_{k,i}, \forall k,i$. Thus, (19) is equivalent to

$$V_1(\tilde{\boldsymbol{r}},\tilde{\boldsymbol{t}},\boldsymbol{S}) \triangleq \sum_{k \in \mathcal{K}}\sum_{i \in \mathcal{K}}(\tilde{r}_{k,i}\tilde{t}_{k,i} - S_{k,i}) \le 0. \tag{20}$$

Now, in order to handle the binary constraint (1), we note that $x \in \{0,1\} \Leftrightarrow x \in [0,1] \ \& \ x - x^2 \le 0$ [9]. Thus, (1) can be replaced by the following equivalent constraints:

$$V_2(\boldsymbol{a}) \triangleq \sum_{k \in \mathcal{N}}\sum_{i \in \mathcal{K}}(a_{k,i} - a_{k,i}^2) \le 0 \tag{21}$$

$$0 \le a_{k,i} \le 1, \forall k,i. \tag{22}$$

Then, problem (10) is written into a more tractable form as

$$\min_{\tilde{\boldsymbol{x}}} q \tag{23a}$$

s.t. $(2),(3),(5),(8),(9),(10b),(11),(14)-(18),(20)-(22)$

$$\sum_{i \in \mathcal{K}}\tilde{t}_{k,i} \le q, \forall k,i, \tag{23b}$$

where $\tilde{\boldsymbol{x}} \triangleq \{\boldsymbol{x},\hat{\boldsymbol{r}},\tilde{\boldsymbol{r}},\hat{\boldsymbol{t}},\tilde{\boldsymbol{t}},q\}$, and q is an additional variable. Let $\mathcal{F} \triangleq \{(2),(3),(5),(8),(9),(10b),(11),(14)-(18),(20)-(22),(23b)\}$ be the feasible set of problem (23). We consider the problem

$$\min_{\tilde{\boldsymbol{x}} \in \widehat{\mathcal{F}}} \mathcal{L}(\boldsymbol{a},\tilde{\boldsymbol{r}},\tilde{\boldsymbol{t}},\boldsymbol{S},\lambda), \tag{24}$$

where $\mathcal{L}(\boldsymbol{a},\tilde{\boldsymbol{r}},\tilde{\boldsymbol{t}},\boldsymbol{S},\lambda) \triangleq q + \lambda(\gamma_1 V_1(\tilde{\boldsymbol{r}},\tilde{\boldsymbol{t}},\boldsymbol{S}) + \gamma_2 V_2(\boldsymbol{a}))$ is the Lagrangian of (23), $\gamma_1, \gamma_2 > 0$ are fixed weights, and λ is the Lagrangian multiplier corresponding to constraints (20), (21). Here, $\widehat{\mathcal{F}} \triangleq \mathcal{F} \backslash \{(20),(21)\}$.

Proposition 1. *The values $V_{1,\lambda}$ and $V_{2,\lambda}$ of V_1 and V_2 at the solution of (24) corresponding to λ converge to 0 as $\lambda \to +\infty$. Moreover, problem (23) has strong duality, i.e.,*

$$\min_{\tilde{\boldsymbol{x}} \in \mathcal{F}} q = \sup_{\lambda \ge 0} \min_{\tilde{\boldsymbol{x}} \in \widehat{\mathcal{F}}} \mathcal{L}(\boldsymbol{a},\tilde{\boldsymbol{r}},\tilde{\boldsymbol{t}},\boldsymbol{S},\lambda), \tag{25}$$

and consequently, (23) is equivalent to (24) at the optimal solution $\lambda^* \ge 0$ of the sup-min problem in (25).

Proof. See Appendix. □

Theoretically, it is required to have $V_{1,\lambda} = 0$ and $V_{2,\lambda} = 0$ in order to obtain the optimal solution to problem (23). By Proposition 1, $V_{1,\lambda}$ and $V_{2,\lambda}$ converge to 0 as $\lambda \to +\infty$. In practice, it is sufficient to accept $V_{1,\lambda} \le \varepsilon, V_{2,\lambda} \le \varepsilon$ for some

---

Algorithm 1 Solving problem (24)

1: **Initialize**: Set $n=0$ and choose a random point $\tilde{\boldsymbol{x}}^{(0)} \in \widehat{\mathcal{F}}$.
2: **repeat**
3:     Update $n = n+1$
4:     Solve (31) to obtain its optimal solution $\tilde{\boldsymbol{x}}^*$
5:     Update $\tilde{\boldsymbol{x}}^{(n)} = \tilde{\boldsymbol{x}}^*$
6: **until** convergence

---

small ε with a sufficiently large value of λ. In our numerical experiments, for $\varepsilon = 10^{-3}$, we see that $\lambda = 1$ with $\gamma_1 = 0.1, \gamma_2 = 0.01$ is enough to ensure that $V_{1,\lambda} \le \varepsilon, V_{2,\lambda} \le \varepsilon$. This way of choosing λ has been widely used in the literature, e.g., see [9] and references therein.

Problem (24) is still difficult to solve due to the nonconvex constraints (14)–(18), and nonconvex parts $V_1(\boldsymbol{a}), V_2(\tilde{\boldsymbol{r}},\tilde{\boldsymbol{t}},\boldsymbol{S})$ in the cost function $\mathcal{L}(\boldsymbol{a},\tilde{\boldsymbol{r}},\tilde{\boldsymbol{t}},\boldsymbol{S},\lambda)$. To deal with (14), we observe that $\log\left(1+\frac{x}{y}\right) \ge \log\left(1+\frac{x^{(n)}}{y^{(n)}}\right) + \frac{2x^{(n)}}{(x^{(n)}+y^{(n)})} - \frac{(x^{(n)})^2}{(x^{(n)}+y^{(n)})x} - \frac{x^{(n)}y}{(x^{(n)}+y^{(n)})y^{(n)}}$, where $x > 0, y > 0$ [10, Eq. (76)]. Therefore, the concave lower bound $\widehat{R}_{d,k,i}(\boldsymbol{\eta}_i)$ of $R_{d,k,i}(\boldsymbol{\eta}_i)$ is given by $\widehat{R}_{d,k,i} \triangleq \frac{\tau_c - \tau_p}{\tau_c \log 2}B\Big[\log\left(1+\frac{\Upsilon_i^{(n)}}{\Phi_i^{(n)}}\right) + \frac{2\Upsilon_i^{(n)}}{(\Upsilon_i^{(n)}+\Phi_i^{(n)})} - \frac{(\Upsilon_i^{(n)})^2}{(\Upsilon_i^{(n)}+\Phi_i^{(n)})\Upsilon_i} - \frac{\Upsilon_i^{(n)}\Phi}{(\Upsilon_i^{(n)}+\Phi_i^{(n)})\Phi_i^{(n)}}\Big]$, where $\Upsilon_i(\eta_{k,i}) \triangleq (M-K_i)\rho\hat{\sigma}_k^2\eta_{k,i}$ and $\Phi_i(\boldsymbol{\eta}_i) \triangleq \rho(\beta_k - \sigma_k^2)\sum_{\ell \in \mathcal{K}}\eta_{\ell,i} + 1$. Then (14) can be approximated by the following convex constraint

$$\hat{r}_{k,i} \le \widehat{R}_{k,i}(\boldsymbol{\eta}_i), \forall k,i. \tag{26}$$

To deal with constraints (15), we observe that $\log\left(1+\frac{x}{y}\right) \le \log\left(x^{(n)}+y^{(n)}\right) + \frac{x+y-x^{(n)}-y^{(n)}}{x^{(n)}+y^{(n)}} - \log(y)$, where $x > 0, y > 0$. Therefore, the convex upper bound $\widetilde{R}_{k,i}(\boldsymbol{\eta}_i)$ of $R_{k,i}(\boldsymbol{\eta}_i)$ is expressed as $\widetilde{R}_{d,k,i} \triangleq \frac{\tau_c - \tau_{u,p}}{\tau_c \log 2}B\Big[\log\left(\Upsilon_i^{(n)}+\Phi_i^{(n)}\right) + \frac{\Upsilon_i+\Phi_i-\Upsilon_i^{(n)}-\Phi_i^{(n)}}{\Upsilon_i^{(n)}+\Phi_i^{(n)}} - \log(\Phi_i)\Big]$. Thus, constraint (15) can be approximated by the following convex constraint

$$\tilde{r}_{k,i} \ge \widetilde{R}_{k,i}(\boldsymbol{\eta}_i), \forall k,i. \tag{27}$$

Next, we observe that $xy \le 0.25[(x+y)^2 - 2(x^{(n)}-y^{(n)})(x-y) + (x^{(n)}-y^{(n)})^2]$ and $-xy \le 0.25[(x-y)^2 - 2(x^{(n)}+y^{(n)})(x+y) + (x^{(n)}+y^{(n)})^2], \forall x \ge 0, y \ge 0, z \ge 0$ [3]. Therefore, (16)–(18) can be approximated by the following convex constraints

$$\hat{t}_{d,k,i} + 0.25[(a_{k,i}-t_{d,i})^2 - 2(a_{k,i}^{(n)}+t_{d,i}^{(n)})(a_{k,i}+t_{d,i}) + (a_{k,i}^{(n)}+t_{d,i}^{(n)})^2] \le 0, \forall k,i \tag{28}$$

$$0.25[(a_{k,i}+t_{d,i})^2 - 2(a_{k,i}^{(n)}-t_{d,i}^{(n)})(a_{k,i}-t_{d,i}) + (a_{k,i}^{(n)}-t_{d,i}^{(n)})^2] - \tilde{t}_{d,k,i} \le 0, \forall k,i \tag{29}$$

$$S_{d,k,i} + 0.25[(\hat{r}_{d,k,i}-\hat{t}_{d,k,i})^2 - 2(\hat{r}_{d,k,i}^{(n)}+\hat{t}_{d,k,i}^{(n)})(\hat{r}_{d,k,i}+\hat{t}_{d,k,i}) + (\hat{r}_{d,k,i}^{(n)}+\hat{t}_{d,k,i}^{(n)})^2] \le 0, \forall k,i. \tag{30}$$

Similarly, the convex upper bounds $\widetilde{V}_1(\boldsymbol{a}), \widetilde{V}_2(\tilde{\boldsymbol{r}},\tilde{\boldsymbol{t}},\boldsymbol{S})$ of the nonconvex parts $V_1(\boldsymbol{a}), V_2(\tilde{\boldsymbol{r}},\tilde{\boldsymbol{t}},\boldsymbol{S})$ are respectively given by

$$\widetilde{V}_1(\tilde{\boldsymbol{r}},\tilde{\boldsymbol{t}},\boldsymbol{S}) \triangleq \sum_{i \in \mathcal{K}}\sum_{k \in \mathcal{N}}0.25[(\tilde{r}_{d,k,i}+\tilde{t}_{d,k,i})^2$$
$$-2(\tilde{r}_{d,k,i}^{(n)}-\tilde{t}_{d,k,i}^{(n)})(\tilde{r}_{d,k,i}-\tilde{t}_{d,k,i}) + (\tilde{r}_{d,k,i}^{(n)}-\tilde{t}_{d,k,i}^{(n)})^2 - 4S_{d,k,i}]$$

$$\widetilde{V}_2(\boldsymbol{a}) \triangleq \sum_{i \in \mathcal{K}}\sum_{k \in \mathcal{N}}(a_{k,i} - 2a_{k,i}^{(n)}a_{k,i} + (a_{k,i}^{(n)})^2).$$

At iteration $(n+1)$, for a given point $\tilde{\boldsymbol{x}}^{(n)}$, problem (24) can finally be approximated by the following convex problem

$$\min_{\tilde{\boldsymbol{x}} \in \widehat{\mathcal{F}}} \widehat{\mathcal{L}}(\boldsymbol{a},\tilde{\boldsymbol{r}},\tilde{\boldsymbol{t}},\boldsymbol{S},\lambda) \tag{31}$$

where $\widehat{\mathcal{L}}(\boldsymbol{a}, \tilde{\boldsymbol{r}}, \tilde{\boldsymbol{t}}, \boldsymbol{S}, \lambda) \triangleq q + \lambda(\gamma_1 \widetilde{V}_1(\boldsymbol{a}) + \gamma_2 \widetilde{V}_2(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{t}}, \boldsymbol{S})$ and $\widetilde{\mathcal{F}} \triangleq \{\{(2), (3), (5), (10b), (11), (22), (23b), (26), (27) - (30)\}$ is a convex feasible set. In Alg. 1, we outline the main steps to solve problem (24). Starting from a random point $\widetilde{\boldsymbol{x}} \in \widehat{\mathcal{F}}$, we solve (31) to obtain its optimal solution $\widetilde{\boldsymbol{x}}^*$, and use $\widetilde{\boldsymbol{x}}^*$ as an initial point in the next iteration. The algorithm terminates when an accuracy level of ε is reached. Alg. 1 will converge to a Fritz John solution of problem (24) (hence (23) or (10)). The proof of this fact is rather standard, and follows from [9, Proposition 2].

Note that Problem (10) is constructed using the achievable rates (5) that depend only on the large-scale coefficients β_k. Before the downlink transmission, the BS solves (10) to obtain the session durations, per-session user assignments, data rates, and transmit powers. Therefore, no extra signalling overhead to schedule UEs/rates and no optimization algorithm are required during the transmission.

B. Conventional Schemes

The conventional schemes can be considered as special cases of the session-based scheme with only one session. Thus, all variables in the conventional schemes can be directly obtained from the session-based scheme by dropping the index i. More precisely, the power constraint at the BS and the achievable rate of UE k in the conventional schemes are, respectively, given by

$$\sum_{k \in \mathcal{K}} \eta_k \leq 1 \qquad (32)$$

and $R_k(\boldsymbol{\eta}) = \frac{\tau_c - \tau_p}{\tau_c} B \log_2 (1 + \mathrm{SINR}_k(\boldsymbol{\eta}))$, where $\boldsymbol{\eta} \triangleq \{\eta_k\}_{k \in \mathcal{K}}$ are power control coefficients, and $\mathrm{SINR}_k(\boldsymbol{\eta}) = \frac{(M-K)\rho \sigma_k^2 \eta_k}{\rho(\beta_k - \sigma_k^2) \sum_{\ell \in \mathcal{K}} \eta_\ell + 1}$. Note that since T_c is in order of milliseconds, and the conventional schemes have only one session, the completion times of UEs is normally larger than T_c, which is confirmed in the numerical results in Section V.

1) Conventional Data Size-Aware Scheme: The corresponding problem of completion time minimization is

$$\min_{\boldsymbol{\eta}} \ \max_k \frac{S_k}{R_k(\boldsymbol{\eta})} \qquad (33a)$$

$$\text{s.t. } (32)$$

$$0 \leq \eta_k, \forall k, \qquad (33b)$$

$$S_k / R_k(\boldsymbol{\eta}) \leq \widetilde{T}_c. \qquad (33c)$$

Problem (33) can be transformed into epigraph form as

$$\min_{\boldsymbol{y}} \ z \qquad (34a)$$

$$\text{s.t. } (32), (33b)$$

$$S_k / r_k \leq z \leq \widetilde{T}_c, \forall k \qquad (34b)$$

$$r_k \leq R_k(\boldsymbol{\eta}), \forall k, \qquad (34c)$$

where $\boldsymbol{y} \triangleq \{\boldsymbol{\eta}, \boldsymbol{r}, z\}$, $\boldsymbol{r} \triangleq \{r_k\}, \forall k$, and z are additional variables. Using the same approach to deal with constraint (14), we obtain the concave lower bound $\widehat{R}_k(\boldsymbol{\eta})$ of $R_k(\boldsymbol{\eta})$ as $\widehat{R}_k \triangleq \frac{\tau_c - \tau_{d,p}}{\tau_c \log 2} B \big[\log \big(1 + \frac{\Upsilon^{(n)}}{\Phi^{(n)}}\big) + \frac{2\Upsilon^{(n)}}{(\Upsilon^{(n)} + \Phi^{(n)})} - \frac{(\Upsilon^{(n)})^2}{(\Upsilon^{(n)} + \Phi^{(n)})\Upsilon} - \frac{\Upsilon^{(n)} \Phi}{(\Upsilon^{(n)} + \Phi^{(n)})\Phi^{(n)}} \big]$, where $\Upsilon(\eta_k) \triangleq (M - K)\rho \sigma_k^2 \eta_k$, $\Phi(\boldsymbol{\eta}) \triangleq \rho(\beta_k - \sigma_k^2) \sum_{\ell \in \mathcal{K}} \eta_\ell + 1$. Then, constraints (34c) can be approximated by the following convex constraint

$$r_k \leq \widehat{R}_k(\boldsymbol{\eta}), \forall k. \qquad (35)$$

Now, at iteration $(n+1)$, for a given point $\boldsymbol{y}^{(n)}$, problem (34) can be approximated by the following convex problem:

$$\min_{\boldsymbol{y} \in \widetilde{\mathcal{H}}} z, \qquad (36)$$

Algorithm 2 Solving problem (34)

1: **Initialize**: Set $n = 0$ and choose a random point $\boldsymbol{y}^{(0)} \in \mathcal{H}$.
2: **repeat**
3:     Update $n = n + 1$
4:     Solve (36) to obtain its optimal solution $\boldsymbol{y}^*$
5:     Update $\boldsymbol{y}^{(n)} = \boldsymbol{y}^*$
6: **until** convergence

---

where $\widetilde{\mathcal{H}} \triangleq \{(32), (33b), (34b), (35)\}$ is a convex feasible set. In Alg. 2, we outline the main steps to solve problem (34). Let $\mathcal{H} \triangleq \{(32), (33b), (34b), (34c)\}$ be the feasible set of problem (34). Starting from a random point $\boldsymbol{y} \in \mathcal{H}$, we solve (36) to obtain its optimal solution $\boldsymbol{y}^*$, and use $\boldsymbol{y}^*$ as an initial point in the next iteration. The algorithm terminates when an accuracy level of ε is reached. Since $\widetilde{\mathcal{H}}$ satisfies Slater's constraint qualification condition, Alg. 2 converges to a Karush–Kuhn–Tucker solution of (34) (hence (33)) [11, Theorem 1].

2) Conventional Non-Data Size-Aware Scheme: This scheme does not take into account the size of the UE data. The completion time is reduced by improving the rates of all UEs. To this end, the scheme aims to maximize the lowest rate of all UEs, which leads to the following problem

$$\max_{\boldsymbol{\eta}} \ \min_k R_k(\boldsymbol{\eta}) \qquad (37a)$$

$$\text{s.t. } (32), (33b).$$

The optimal solution to problem (37) can be written in closed-form as $\eta_k = \frac{1 + \rho(\beta_k - \sigma_k^2)}{\rho \sigma_k^2 \left(\frac{1}{\rho} \sum_{\ell \in \mathcal{K}} \frac{1}{\sigma_\ell^2} + \sum_{\ell \in \mathcal{K}} \frac{\beta_\ell - \sigma_\ell^2}{\sigma_\ell^2}\right)}$ [5, Tab. 5.4].

3) Heuristic Small-Scale-Fading-Based Scheme: In each small-scale coherence block, greedy user scheduling and power allocation to (approximately) maximize the lowest rate are performed. Specifically, if the data queue of a UE becomes zero, this UE will be no longer scheduled in the subsequent small-scale coherence block. This way, the remaining UEs at later small-scale coherence blocks will have more power and higher transmission rate, which eventually contributes to reducing the longest completion time. The optimal power control for the max-min rate problem is $\eta_k = 1/c_k / (\sum_{k \in \mathcal{K}} (1/c_k))$, where $c_k = \rho |\boldsymbol{g}_k^T \boldsymbol{u}_k|^2$.

V. NUMERICAL RESULTS

We consider a square-shaped cell of size $D \times D$, where $D = 0.25$ km. The BS is at the center, and the UEs are randomly located. We set $\tau_c = 200$ samples. The large-scale fading coefficients, β_k, are modeled as in [12], $\beta_k[\mathrm{dB}] = -148.1 - 37.6 \log_{10} \left(\frac{d_k}{1 \text{ km}}\right) + z_k$, where $d_k \geq 35$ m is the distance between UE k and the BS, and z_k represents shadow fading, which has zero mean and 7 dB standard deviation. We take the bandwidth to $B = 100$ MHz and the noise power to $\sigma_0^2 = -92$ dBm. Let $\tilde{\rho} = 1$ W and $\tilde{\rho}_p = 0.1$ W be the maximum transmit power of the BS and uplink pilot sequences, respectively. The maximum transmit powers ρ and ρ_p are normalized by the noise power. The sizes of the UE data are taken to monotonically increase from UE 1 to UE K with a step Δ, i.e., $S_{k+1} = S_k + \Delta$. Here, $S_1 = 0.125$ MB and $\Delta = 0.5$ MB. We set $T_c = 1$ms and $\widetilde{T}_c = 10$s.

Fig. 2 compares the completion time per UE of the proposed session-based (**SB**) scheme to those of the conventional data size-aware (**ConDA**), non-data size-aware (**ConNoDA**) schemes, and heuristic scheme on the small-scale fading time scale (**SmallScale**). The results in Fig. 2 are obtained using 200
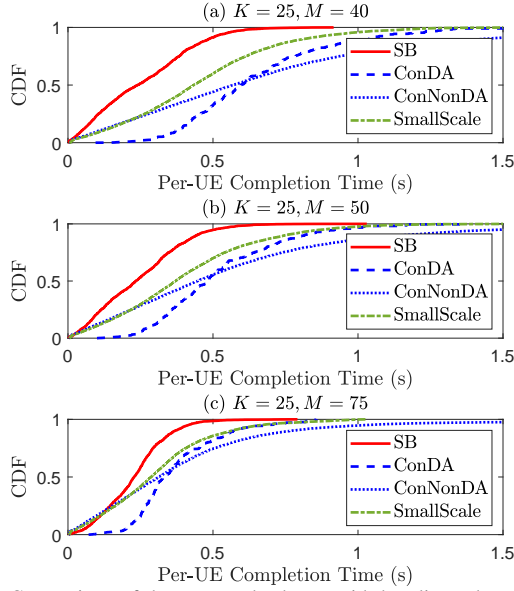
Fig. 2. Comparison of the proposed scheme with baseline schemes.

channels. The maximum completion time of UEs of **ConNoDA** is smaller than $\widetilde{T}_c$ and the minimum is larger than T_c. As seen, in terms of 90%-likely performance, the session-based scheme significantly outperforms **ConDA**, **ConNoDA**, and **SmallScale** when M/K is small. Specifically, the 90%-likely completion time per UE of **SB** with $K = 25, M = 40$ is 0.48s, which is $2 - 3$ times smaller than those of **SmallScale**, **ConDA** and **ConNoDA**. For the case of $(K = 25, M = 75)$ where M/K is already large, the inter-user interference is relatively small, and hence, the improvement by our session-based scheme becomes smaller. The baseline schemes can outperform the proposed scheme in some cases but only in terms of less-than-25%-likely performance. Fig. 2 also shows the advantage of joint optimization of user assignment, time, and rates over the small-scale allocation approach.

VI. CONCLUSION

In this work, we have proposed a session-based scheme for the massive MIMO downlink, where the BS knows, a priori, the amount of data sent to the UEs. We formulated an optimization problem for assigning UEs to sessions and allocating power to minimize the completion time of the UEs. Utilizing successive convex approximation techniques, we proposed a novel algorithm to solve the formulated problem. Numerical results showed that our session-based scheme can significantly reduce the completion time compared with conventional schemes.

APPENDIX

The proof basically follows the arguments in [9] with some modifications for our setting. Denote by $\mathcal{L}_\lambda$ the optimal value of problem (24) corresponding to λ. Also, denote by q^* the optimal value of problem (23). Then $q^* < +\infty$ due to the compactness of $\mathcal{F}$. By a duality gap between the optimal values of problem (23) and its dual problem,

$$\sup_{\lambda \geq 0} \mathcal{L}_\lambda = \sup_{\lambda \geq 0} \min_{\widetilde{\boldsymbol{x}} \in \widehat{\mathcal{F}}} \mathcal{L}(\boldsymbol{a}, \tilde{\boldsymbol{r}}, \tilde{\boldsymbol{t}}, \boldsymbol{S}) \leq q^* = \min_{\widetilde{\boldsymbol{x}} \in \widehat{\mathcal{F}}} \max_{\lambda \geq 0} \mathcal{L}(\boldsymbol{a}, \tilde{\boldsymbol{r}}, \tilde{\boldsymbol{t}}, \boldsymbol{S}).$$

It follows that, for all $\lambda \geq 0$,

$$\mathcal{L}_\lambda \leq q^* < +\infty. \tag{38}$$

For each $\lambda \geq 0$, let $V_{1,\lambda} \triangleq \sum_{k \in \mathcal{K}} \sum_{i \in \mathcal{K}} ((\tilde{r}_{k,i})_\lambda (\tilde{t}_{k,i})_\lambda - (S_{k,i})_\lambda)$ and $V_{2,\lambda} \triangleq \sum_{k \in \mathcal{N}} \sum_{i \in \mathcal{K}} ((a_{k,i})_\lambda - (a_{k,i})_\lambda^2))$ be the

values of V_1 and V_2 at the optimal solution $(\boldsymbol{a}_\lambda, \tilde{\boldsymbol{r}}_\lambda, \tilde{\boldsymbol{t}}_\lambda, \boldsymbol{S}_\lambda)$ of (24) corresponding to λ. We see from (14)–(18) that $V_{1,\lambda} \geq 0$ and from (22) that $V_{2,\lambda} \geq 0$. Set $V_\lambda \triangleq \gamma_1 V_{1,\lambda} + \gamma_2 V_{2,\lambda}$ and let q_λ be the value of q corresponding to λ. Next, let $0 \leq \lambda_1 < \lambda_2$. By the definition of $\mathcal{L}_{\lambda_1}$ and $\mathcal{L}_{\lambda_2}$,

$$\mathcal{L}_{\lambda_1} = q_{\lambda_1} + \lambda_1 V_{\lambda_1} \leq q_{\lambda_2} + \lambda_1 V_{\lambda_2}, \tag{39}$$

$$\mathcal{L}_{\lambda_2} = q_{\lambda_2} + \lambda_2 V_{\lambda_2} \leq q_{\lambda_1} + \lambda_2 V_{\lambda_1}, \tag{40}$$

from which we have $\lambda_1 V_{\lambda_1} + \lambda_2 V_{\lambda_2} \leq \lambda_1 V_{\lambda_2} + \lambda_2 V_{\lambda_1}$, and so $V_{\lambda_2} \leq V_{\lambda_1}$. This means V_λ is decreasing as λ is increasing. Since $V_\lambda = \gamma_1 V_{1,\lambda} + \gamma_2 V_{2,\lambda} \geq 0$ for all $\lambda \geq 0$, we obtain that $V_\lambda \to V^* \geq 0$ as $\lambda \to +\infty$. From (39) and (40), we also have that $\lambda_2 q_{\lambda_1} + \lambda_1 q_{\lambda_2} \leq \lambda_2 q_{\lambda_2} + \lambda_1 q_{\lambda_1}$, which yields $q_{\lambda_2} \geq q_{\lambda_1}$. Therefore, as $\lambda \to +\infty$, q_λ is increasing and hence bounded from below. Now, if $V^* > 0$, then $\mathcal{L}_\lambda = q_\lambda + \lambda V_\lambda \to +\infty$ as $\lambda \to +\infty$, which contradicts (38). Thus, we must have $V^* = 0$, that is, $V_\lambda \to 0$ as $\lambda \to +\infty$, which implies that $V_{1,\lambda} \to 0$ and $V_{2,\lambda} \to 0$ as $\lambda \to +\infty$.

Finally, let $\widetilde{\boldsymbol{x}}_\lambda$ be the value of $\widetilde{\boldsymbol{x}}$ corresponding to λ. Then $\widetilde{\boldsymbol{x}}_\lambda \in \widehat{\mathcal{F}}$. Since $\widehat{\mathcal{F}}$ is bounded, there exists a cluster point $\widetilde{\boldsymbol{x}}_*$ of $\{\widetilde{\boldsymbol{x}}_\lambda\}_\lambda$ as $\lambda \to +\infty$. We assume, without loss of generality, that $\widetilde{\boldsymbol{x}}_\lambda \to \widetilde{\boldsymbol{x}}_*$. Then, $\boldsymbol{a}_\lambda \to \boldsymbol{a}_*$, $\tilde{\boldsymbol{r}}_\lambda \to \tilde{\boldsymbol{r}}_*$, $\boldsymbol{S}_\lambda \to \boldsymbol{S}_*$, and $\tilde{\boldsymbol{t}}_\lambda \to \tilde{\boldsymbol{t}}_*$. It follows that $V_{1,\lambda} \to (V_1)_* \triangleq \sum_{k \in \mathcal{K}} \sum_{i \in \mathcal{K}} ((\tilde{r}_{k,i})_* (\tilde{t}_{k,i})_* - (S_{k,i})_*)$, $V_{2,\lambda} \to (V_2)_* \triangleq \sum_{k \in \mathcal{N}} \sum_{i \in \mathcal{K}} ((a_{k,i})_* - (a_{k,i})_*^2))$, $V_\lambda \to V_* \triangleq \gamma_1 (V_1)_* + \gamma_2 (V_2)_*$, and $q_\lambda \to q_*$. As shown above, $(V_1)_* = 0$, $(V_2)_* = 0$, and $V_* = 0$. Therefore, $(\tilde{\boldsymbol{r}}_*, \tilde{\boldsymbol{t}}_*, \boldsymbol{S}_*)$ and $(\boldsymbol{a}_*)$ satisfy (20) and (21), respectively. This together with $\widetilde{\boldsymbol{x}}_* \in \widehat{\mathcal{F}}$ implies that $\widetilde{\boldsymbol{x}}_* \in \mathcal{F}$, and so $\widetilde{\boldsymbol{x}}_*$ is a feasible point of (23). As such, $q_* \geq q^*$. Next, the definition of $\mathcal{L}_\lambda$ implies that, for all $\lambda \geq 0$, $\sup_{\lambda \geq 0} \mathcal{L}_\lambda \geq \mathcal{L}_\lambda = q_\lambda + \lambda V_\lambda \geq q_\lambda$. By letting $\lambda \to +\infty$, $\sup_{\lambda \geq 0} \mathcal{L}_\lambda \geq q_* \geq q^*$. Combining with (38), yields $\sup_{\lambda \geq 0} \widetilde{\mathcal{L}}(\lambda) = q_* = q^*$. We conclude that (25) holds and that $\widetilde{\boldsymbol{x}}_*$ is an optimal solution of (23), which completes the proof.

REFERENCES

[1] N. Alexander *et al.*, "The future of livestreaming: Strategy and predictive analysis for future operations of facebook live," in *Prof. IEEE SIEDS*, Apr. 2021, pp. 1–6.

[2] L. De Cicco and S. Mascolo, "An adaptive video streaming control system: Modeling, validation, and performance evaluation," *IEEE/ACM Transactions on Networking*, vol. 22, no. 2, pp. 526–539, Apr. 2014.

[3] T. T. Vu *et al.*, "Cell-free massive MIMO for wireless federated learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6377–6392, Oct. 2020.

[4] H. Yan *et al.*, "Discovering usage patterns of mobile video service in the cellular networks," *IEEE Trans. Netw. Serv. Manag.*, vol. 18, no. 2, pp. 1789–1802, Jun. 2021.

[5] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Cambridge University Press, 2016.

[6] Y. Liu and E. Erkip, "Completion time in two-user channels: An information-theoretic perspective," *IEEE Trans. Inf. Theory*, vol. 63, no. 5, pp. 3209–3223, May 2017.

[7] T. T. Vu *et al.*, "Energy-efficient massive MIMO for federated learning: Transmission designs and resource allocations," Dec. 2021. [Online]. Available: https://arxiv.org/abs/2112.11723

[8] M. Feng, M. Krunz, and W. Zhang, "Joint task partitioning and user association for latency minimization in mobile edge computing networks," *IEEE Trans. Veh. Technol.*, vol. 70, no. 8, pp. 8108–8121, Aug. 2021.

[9] T. T. Vu *et al.*, "Spectral and energy efficiency maximization for content-centric C-RANs with edge caching," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6628–6642, Dec. 2018.

[10] L. D. Nguyen *et al.*, "Energy-efficient multi-cell massive MIMO subject to minimum user-rate constraints," *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 914–928, Feb. 2021.

[11] B. R. Marks and G. P. Wright, "A general inner approximation algorithm for nonconvex mathematical programs," *Operations Research*, vol. 26, no. 4, pp. 681–683, Aug. 1978.

[12] *Further Advancements for E-UTRA Physical Layer Aspects (Release 9)*, document TS 36.814, 3GPP, Mar. 2010.