# Low Latency Driven Effective Capacity Analysis for Non-orthogonal and Orthogonal Spectrum Access

Wenjuan Yu[1], Leila Musavian[2], Atta ul Quddus[1], Qiang Ni[3] and Pei Xiao[1]
[1]5GIC, Institute of Communication Systems, University of Surrey, Guildford, GU2 7XH, UK
[2]School of Computer Science and Electronic Engineering, University of Essex, UK
[3]School of Computing and Communications, InfoLab21, Lancaster University, UK
Emails: {w.yu, a.quddus, p.xiao}@surrey.ac.uk, leila.musavian@essex.ac.uk, q.ni@lancaster.ac.uk

***Abstract*—In this paper, we theoretically investigate the performance of non-orthogonal and orthogonal spectrum access protocols (more generically known as NOMA) in supporting ultra-reliable low-latency communications (URLLC). The theory of effective capacity (EC) is adopted as a suitable delay-guaranteed capacity metric to flexibly represent the users' delay requirements. Then, the total EC difference between a downlink user-paired NOMA network and a downlink orthogonal multiple access (OMA) network is analytically studied. Exact closed-form expressions and the approximated closed-forms at high signal-to-noise ratios (SNRs) are derived for both networks and validated through simulation results. It is shown that for a user pair in which two users with the most distinct channel conditions are paired together, NOMA still achieves higher total EC (compared to OMA) in high SNR regime as the user group size becomes larger, although the EC performance of both NOMA and OMA reduces with the increase in group size. It is expected that the derived analytical framework can serve as a useful reference and practical guideline for designing favourable orthogonal and non-orthogonal spectrum access schemes in supporting low-latency services.**

***Index Terms*—URLLC, NOMA, effective capacity, delay violation probability, exact closed-form expressions.**

## I. Introduction

With the explosive growth of delay-sensitive wireless communication applications and services such as vehicular communications, tactile Internet and virtual reality, the ability of supporting low-latency services becomes more and more important. The provision of ultra-reliable low-latency communications (URLLC) has also emphasized the importance of reliability and low-latency transmissions in the 5th generation (5G) cellular networks and beyond [2]. In this paper, we adopt the concept of effective capacity (EC) to describe the maximum constant arrival rate that can be served, while guaranteeing a statistical delay requirement [3]. From the introduction about EC given in Section III, it will become evident that EC is not only a suitable, but also a flexible, metric which can represent various delay requirements. This allows us to investigate the delay-constrained data rates that can be served by a proposed communication network, for different latency requirements.

Furthermore, due to the spectrum scarcity issue in future communication networks, it is important to study and investigate the performance of spectrum-efficient multiple access protocols. In recent years, non-orthogonal multiple access (NOMA)[1] was introduced as a promising multiple access (MA) technique for 5G communications and it has been shown to be spectrally efficient since it allows multiple users to transmit with different power levels, but using the same subcarriers and time slots. Further, NOMA has been shown to have some advantages over conventional orthogonal multiple access (OMA) in different communication systems [4]–[9]. For a downlink transmission with NOMA applied, the base station (BS) will send a superimposed mixture containing all users' signals, then the users with stronger channel conditions can obtain the information of the users with weaker channel conditions in accordance with NOMA principle. This can be exploited to improve the weaker user's reception reliability [7]. Moreover, a NOMA network with fixed power allocation (F-NOMA) was compared with the conventional OMA in [5], which shows that F-NOMA can achieve a larger sum rate than OMA and if two users with very different channel conditions are paired, the performance of NOMA can be further improved. In [7], a non-orthogonal relaying scheme was proposed and studied, which improves the ergodic secrecy rate, compared to the conventional orthogonal relaying schemes.

Although the aforementioned studies were conducted to explore the applications of NOMA in different communication networks, they are more suitable for delay-insensitive services [10]. Focusing on delay-sensitive applications, the sum EC was formulated and maximized for a downlink NOMA network in [11], which was then solved with a suboptimal power control policy. In [12], the max-min EC problem was first proved to be quasi-concave and then solved using a bisection-based power allocation algorithm. However, [11] and [12] focused on designing efficient power allocation algorithms, rather than analytically analyzing the performance of NOMA in supporting delay-sensitive services. Considering the users with and without delay requirements, the exact analytical closed-form expressions for the ECs were proposed in [1] for a two-user NOMA network only. However, when *multiple* NOMA pairs are considered, the exact and the approximated closed-form expressions at high transmit signal-to-noise ratios (SNRs) for EC are not provided in [1].

Part of this conference paper was presented in [1].

[1]We only consider power-domain NOMA in this work.

To investigate orthogonal and non-orthogonal multiple access protocols for low-latency services, in this paper we conduct a comparative study on a downlink user-paired NOMA network and a classical OMA network. The derived analytical framework can serve as a useful reference and practical guideline for designing favourable spectrum access protocols in supporting delay-sensitive communication scenarios. Specifically, we summarize the primary contributions as follows:

- For a downlink NOMA network with multiple pairs and a downlink OMA network, the exact analytical closed-form expressions are derived respectively, followed by the approximated closed-forms for EC at high SNRs. The accuracy of these obtained closed-form expressions is validated through simulation results.
- In Lemma 1, we prove that the performance gain of NOMA over OMA, in supporting delay-sensitive services, becomes stable in the high SNR regime. Simulation results confirm the theoretical analysis and further reveal the impact of delay requirements, power coefficient settings and user group size on the total EC difference between NOMA and OMA schemes.

## II. System Models

Consider a classical downlink transmission with one BS and $K$ mobile users with single antenna employed. Without loss of generality, the $K$ users' channels are assumed to be ordered as $|h_1|^2 \leq |h_2|^2 \leq \cdots \leq |h_K|^2$, where $h_k$ is the channel gain from the BS to the $k^{\text{th}}$ user, following Rayleigh fading distribution with unit variance. Block fading channel models are assumed, i.e., the channel gain stays fixed within one fading-block and independently varies for the next fading-block. The fading-block length, $T_{\text{f}}$, is assumed to be equal to one frame size, which is an integer multiple of the symbol length. Two downlink networks with different multiple access protocols are described as follows.

### A. A Downlink NOMA Network with Multiple User Pairs

For the downlink NOMA network, all $K$ users are grouped into $K/2$ NOMA pairs[2], with the $j^{\text{th}}$ pair described as $\boldsymbol{\phi}_j = \left\{(\mathrm{u}_j, \mathrm{w}_j) \mid \mathrm{u}_j \neq \mathrm{w}_j, |h_{\mathrm{u}_j}|^2 \leq |h_{\mathrm{w}_j}|^2\right\}, \forall j \in \{1, 2, \ldots, K/2\}$. Within each pair, NOMA is applied and for the inter-pair MA, the conventional OMA is utilized. For the $j^{\text{th}}$ pair, the BS will send $\sqrt{\alpha_{\mathrm{u}_j} P} s_{\mathrm{u}_j} + \sqrt{\alpha_{\mathrm{w}_j} P} s_{\mathrm{w}_j}$ to both users on the same subcarrier/time-slot in accordance with the NOMA protocol. Here, $\alpha_{\mathrm{u}_j}$ and $\alpha_{\mathrm{w}_j}$ are the power coefficients for the users $\mathrm{u}_j$ and $\mathrm{w}_j$, $P$ is the total transmission power at the BS, $s_{\mathrm{u}_j}$ and $s_{\mathrm{w}_j}$ are the messages for the users $\mathrm{u}_j$ and $\mathrm{w}_j$. By employing successive interference cancellation (SIC), the stronger user $\mathrm{w}_j$ with higher channel gains can detect and eliminate the message for the weaker user $\mathrm{u}_j$, before it decodes its own. The user $\mathrm{u}_j$ with smaller channel gains will decode its information by considering the user $\mathrm{w}_j$'s message as noise. It is assumed that $R_{\mathrm{u}_j \to \mathrm{w}_j} \geq \tilde{R}_{\mathrm{u}_j}$ [5], so that SIC is guaranteed to be successfully applied at the user $\mathrm{w}_j$. Here, $R_{\mathrm{u}_j \to \mathrm{w}_j}$ denotes the user $\mathrm{w}_j$'s data rate to decode the user $\mathrm{u}_j$'s message and $\tilde{R}_{\mathrm{u}_j}$ is the target data rate for the user $\mathrm{u}_j$. Assume that the target rate $\tilde{R}_{\mathrm{u}_j}$ equals to its achievable rate when the user $\mathrm{u}_j$ decodes its own message. Hence, it can be found that the requirement $R_{\mathrm{u}_j \to \mathrm{w}_j} \geq \tilde{R}_{\mathrm{u}_j}$ always holds since $|h_{\mathrm{u}_j}|^2 \leq |h_{\mathrm{w}_j}|^2$ [5].

Fixed power coefficients are considered for all NOMA pairs in this paper, which are non-adaptive and follow NOMA principle, i.e., $\boldsymbol{\alpha}_j = \left\{(\alpha_{\mathrm{u}_j}, \alpha_{\mathrm{w}_j}) \mid \alpha_{\mathrm{u}_j} \geq \alpha_{\mathrm{w}_j}, \alpha_{\mathrm{u}_j} + \alpha_{\mathrm{w}_j} = 1\right\}$, $\forall j \in \{1, 2, \ldots, K/2\}$. Different non-adaptive transmit power policies will be simulated and discussed in Section IV, for the purpose of providing guidance on designing practical power allocation algorithms in future studies.

Then, for the $j^{\text{th}}$ NOMA pair, the achievable data rates for both users can be respectively given by

$$R_{\mathrm{u}_j} = \frac{2}{K} \log_2 \left(1 + \frac{\rho \alpha_{\mathrm{u}_j} |h_{\mathrm{u}_j}|^2}{\rho \alpha_{\mathrm{w}_j} |h_{\mathrm{u}_j}|^2 + 1}\right), \quad \text{(b/s/Hz)} \tag{1a}$$

$$R_{\mathrm{w}_j} = \frac{2}{K} \log_2 \left(1 + \rho \alpha_{\mathrm{w}_j} |h_{\mathrm{w}_j}|^2\right), \quad \text{(b/s/Hz)} \tag{1b}$$

where $\rho$ is the transmit SNR, i.e., $\rho = \dfrac{P}{N_0 B}$. Here, $N_0 B$ is the noise power with $N_0$ indicating the one-sided noise spectral density and $B$ denoting the channel bandwidth.

### B. A Conventional Downlink OMA Network

For a classical downlink OMA network applying time division multiple access (TDMA), each user is assumed to equally occupy $1/K$ of orthogonal resources, which is a typical allocation strategy. Although in this case, the available radio resources for each user become less, the transmit power for each user can be higher. Then, the achievable rate for the user $k$, in b/s/Hz, is given by

$$\bar{R}_k = \frac{1}{K} \log_2 \left(1 + \rho |h_k|^2\right), \ k \in \{1, 2, \ldots, K\}. \tag{2}$$

## III. Effective Capacity

Assume that there exists a virtual buffer for the $k^{\text{th}}$ user at the BS, with an infinite buffer size. Define $D_k(t)$ as the delay experienced by a packet arriving at time $t$. According to [13], [14], the probability of the delay $D_k(t)$ exceeding a maximum delay bound $D_{\max}^k$ can be estimated as

$$P_{\text{delay}}^{\text{out}} = \Pr\{D_k(t) > D_{\max}^k\} \approx \Pr\{Q(t) > 0\} \mathrm{e}^{-\theta_k \mu D_{\max}^k}, \tag{3}$$

where $P_{\text{delay}}^{\text{out}}$ denotes the delay violation probability limit for the $k^{\text{th}}$ user, $\Pr\{Q(t) > 0\}$ is the probability of a non-empty buffer at time $t$ with $Q(t)$ indicating the buffer length, $D_{\max}^k$ is the given delay bound in the unit of symbol duration, and $\theta_k (\theta_k > 0)$ represents the $k^{\text{th}}$ user's exponential decay rate. It was proved that the constant arrival rate has to be limited to the value of $\mu$, which equals to the EC, so that a target delay violation probability limit $P_{\text{delay}}^{\text{out}}$ can be met.

Assume that the given wireless link can support $C_k(t)$ packets per unit of time, which can be modeled as a stationary and ergodic random process with nonnegative values. Assuming that the service process satisfies Gärtner-Ellis theorem [15], the EC for the $k^{\text{th}}$ user on a block-fading channel is defined as

$$E_{\text{c}}^k = -\frac{1}{\theta_k T_{\text{f}} B} \ln \left(\mathbb{E}\left[e^{-\theta_k T_{\text{f}} B C_k}\right]\right), \quad \text{(b/s/Hz)} \tag{4}$$

[2]Here, it is required that $K$ is an even number.

where $\mathbb{E}[\cdot]$ is the expectation over the user $k$'s channel. Set $x_k = \rho|h_k|^2$. Since all $K$ users' channels are assumed to be already ordered, the probability density function (PDF) of the ordered $x_k$, $\forall k \in \{1,\ldots,K\}$, denoted by $f_{(k)}(x_k)$, follows the theory of order statistics and is given by [16]

$$f_{(k)}(x_k) = \psi_k f(x_k) F(x_k)^{k-1} \left(1 - F(x_k)\right)^{K-k}, \quad (5)$$

where $\psi_k = 1/B(k, K-k+1)$, in which the beta function $B(k, K-k+1)$ equals to $\frac{k!(K-k+1)!}{(K+1)!}$ [17]. In (5), $f(x_k)$ and $F(x_k)$ are the PDF and the cumulative distribution function (CDF) of the unordered random variable $x_k$.

With different multiple access protocols applied, the available service rate for each user can be different. In the following, we study the total maximum achievable delay-constrained rate, i.e., the total EC, for the downlink transmission with NOMA applied and also the conventional OMA principle.

### A. For A Downlink NOMA With Multiple User Pairs

By replacing $C_k$ in (4) with the achievable rates (1a) and (1b), the individual EC for the users $\mathrm{u}_j$ and $\mathrm{w}_j$ in the $j^{\text{th}}$ NOMA pair can be respectively given by

$$E_{\mathrm{c}}^{\mathrm{u}_j} = -\frac{1}{\theta_{\mathrm{u}_j} T_{\mathrm{f}} B} \ln\left(\mathbb{E}\left[\left(\frac{\rho|h_{\mathrm{u}_j}|^2+1}{\rho\alpha_{\mathrm{w}_j}|h_{\mathrm{u}_j}|^2+1}\right)^{\frac{4}{K}\beta_{\mathrm{u}_j}}\right]\right), \quad (6a)$$

$$E_{\mathrm{c}}^{\mathrm{w}_j} = -\frac{1}{\theta_{\mathrm{w}_j} T_{\mathrm{f}} B} \ln\left(\mathbb{E}\left[\left(1+\rho\alpha_{\mathrm{w}_j}|h_{\mathrm{w}_j}|^2\right)^{\frac{4}{K}\beta_{\mathrm{w}_j}}\right]\right), \quad (6b)$$

where $\beta_{\mathrm{u}_j} = -\frac{\theta_{\mathrm{u}_j} T_{\mathrm{f}} B}{2\ln 2}$ and $\beta_{\mathrm{w}_j} = -\frac{\theta_{\mathrm{w}_j} T_{\mathrm{f}} B}{2\ln 2}$. The total EC for multiple NOMA pairs, denoted by $\boldsymbol{E}_{\mathrm{c}}^{N}$, equals to $\sum_{j=1}^{K/2}\left(E_{\mathrm{c}}^{\mathrm{u}_j} + E_{\mathrm{c}}^{\mathrm{w}_j}\right)$. We then provide the exact analytical closed-forms and approximations at high SNRs, for $E_{\mathrm{c}}^{\mathrm{u}_j}$ and $E_{\mathrm{c}}^{\mathrm{w}_j}$ in the downlink transmission with NOMA.

*Theorem 1:* The exact closed-forms for $E_{\mathrm{c}}^{\mathrm{u}_j}$ and $E_{\mathrm{c}}^{\mathrm{w}_j}$ are respectively given in (7a) and (7b) (shown on next page). At high SNRs, $\lim_{\rho\to\infty} E_{\mathrm{c}}^{\mathrm{u}_j} = \frac{2}{K}\log_2\left(\frac{1}{\alpha_{\mathrm{w}_j}}\right)$ and $E_{\mathrm{c}}^{\mathrm{w}_j}$ can be approximated by (7c).

*Proof:* The proof for the exact closed-forms is similar to that of two-user case in [1], so it is omitted here due to page limit. The proof for the approximated closed-forms at high SNRs is given in Appendix. ∎

### B. For A Conventional Downlink OMA Network

By inserting (2) into (4), the individual EC for the user $k$ in a conventional OMA network can be given by

$$\bar{E}_{\mathrm{c}}^{k} = -\frac{1}{\theta_k T_{\mathrm{f}} B} \ln\left(\mathbb{E}\left[\left(1+\rho|h_k|^2\right)^{\frac{2}{K}\beta_k}\right]\right). \quad (8)$$

The total EC in a conventional downlink OMA network, denoted by $\boldsymbol{E}_{\mathrm{c}}^{O}$, equals to $\sum_{k=1}^{K}\bar{E}_{\mathrm{c}}^{k}$. The following two theorems derive the exact analytical closed-forms and also the closed-forms at high SNRs for $\bar{E}_{\mathrm{c}}^{k}$, $\forall k \in \{1,\ldots,K\}$.

*Theorem 2:* The exact closed-form for $\bar{E}_{\mathrm{c}}^{k}$ for the user $k$ is given in (7d). At high SNRs, the approximated closed-form for $\bar{E}_{\mathrm{c}}^{k}$ is given in (7e).

*Proof:* The proof for the exact closed-forms is similar to that of two-user case in [1], hence it is omitted due to page limit. The proof for the approximated closed-form at high SNRs is given in Appendix. ∎

### C. Comparison of Total Effective Capacity

The total EC difference between a downlink user-paired NOMA and a conventional OMA can be expressed as

$$\boldsymbol{E}_{\mathrm{c}}^{N} - \boldsymbol{E}_{\mathrm{c}}^{O} = \sum_{j=1}^{K/2}\left(E_{\mathrm{c}}^{\mathrm{u}_j} + E_{\mathrm{c}}^{\mathrm{w}_j}\right) - \sum_{k=1}^{K}\bar{E}_{\mathrm{c}}^{k}, \quad (9)$$

and its exact closed-form and approximation at high SNRs can be obtained, by simply applying the analytical results given in (7a)-(7e).

*Lemma 1:* At extreme SNR values, $\lim_{\rho\to 0}\boldsymbol{E}_{\mathrm{c}}^{N} - \boldsymbol{E}_{\mathrm{c}}^{O} = 0$ and $\lim_{\rho\to\infty}\boldsymbol{E}_{\mathrm{c}}^{N} - \boldsymbol{E}_{\mathrm{c}}^{O}$ approaches a constant, given in (10).

$$\lim_{\rho\to\infty}\boldsymbol{E}_{\mathrm{c}}^{N} - \boldsymbol{E}_{\mathrm{c}}^{O} = \sum_{j=1}^{K/2} -\frac{1}{\theta_{\mathrm{u}_j} T_{\mathrm{f}} B}\ln\left(\frac{\left(\alpha_{\mathrm{w}_j}\right)^{-\frac{4}{K}\beta_{\mathrm{u}_j}}}{\mathbb{E}\left[\left(|h_{\mathrm{u}_j}|^2\right)^{\frac{2}{K}\beta_{\mathrm{u}_j}}\right]}\right) - \frac{1}{\theta_{\mathrm{w}_j} T_{\mathrm{f}} B}\ln\left(\frac{\left(\alpha_{\mathrm{w}_j}\right)^{\frac{4}{K}\beta_{\mathrm{w}_j}}\mathbb{E}\left[\left(|h_{\mathrm{w}_j}|^2\right)^{\frac{4}{K}\beta_{\mathrm{w}_j}}\right]}{\mathbb{E}\left[\left(|h_{\mathrm{w}_j}|^2\right)^{\frac{2}{K}\beta_{\mathrm{w}_j}}\right]}\right). \quad (10)$$

*Proof:* Please refer to Appendix G in [1]. ∎

Numerical results in Section IV show that the downlink user-paired NOMA supports higher total EC values than the downlink OMA network, in high SNR regime. From Lemma 1, we can then conclude that the performance gain of NOMA over OMA approaches a constant value at very high SNRs.

Furthermore, from (10), we notice that the difference of the total ECs, i.e., $\boldsymbol{E}_{\mathrm{c}}^{N} - \boldsymbol{E}_{\mathrm{c}}^{O}$, depends on the user pairing setting $\boldsymbol{\phi}_j$ and also the power coefficients $\boldsymbol{\alpha}_j$, $\forall j \in \{1,2,\ldots,K/2\}$. To investigate the influence of power coefficients, two different power settings are utilized and simulated in Section IV, in order to provide more comprehensive EC comparison results. The optimal user paring algorithms are out of the scope of this paper, but will be considered as a future research topic.

## IV. Numerical Results

In this section, we first validate the accuracy of the proposed exact analytical closed-forms and the approximated closed-forms at high SNRs for $E_{\mathrm{c}}^{\mathrm{u}_j}$, $E_{\mathrm{c}}^{\mathrm{w}_j}$ and $\bar{E}_{\mathrm{c}}^{k}$. The theoretical insights given in Lemma 1 will be numerically confirmed as well. Then, the total EC comparison results between the multiple NOMA pairs and the downlink OMA network will be numerically studied. In our simulations, it is assumed that there are 6 mobile users, i.e, $K = 6$, and 3 NOMA pairs in total. Two different power coefficients settings are utilized: 1) a fixed power allocation; 2) a varied power setting according to [9], i.e., $\alpha_i = \frac{\mathcal{M}-i+1}{\delta}$ for all $\mathcal{M}$ users sharing one radio resource. Here, $\delta$ is to ensure $\sum_{i=1}^{\mathcal{M}}\alpha_i = 1$. The delay exponents are assumed to be the same for all users and are set as $0.01$, unless otherwise indicated. Further, it is assumed that $T_{\mathrm{f}} = 0.5$ ms and $B = 180$kHz.

Assume that the $1^{\text{st}}$ user and the $4^{\text{th}}$ user are paired together in the downlink NOMA network, i.e., $\mathrm{u}_j = 1$ and $\mathrm{w}_j = 4$. The power coefficients for this pair are given as $\alpha_{\mathrm{u}_j} = 0.8$

$$E_{\mathrm{c}}^{\mathrm{u}_j}=-\frac{1}{\theta_{\mathrm{u}_j}T_{\mathrm{f}}B}\ln\Bigg(\frac{(\alpha_{\mathrm{w}_j})^{-\frac{4}{K}\beta_{\mathrm{u}_j}}\psi_{\mathrm{u}_j}}{\rho}\Bigg(\sum_{s=0}^{\mathrm{u}_j-1}\binom{\mathrm{u}_j-1}{s}(-1)^s\frac{\rho}{K-\mathrm{u}_j+1+s}-\frac{4}{K}\beta_{\mathrm{u}_j}\frac{\alpha_{\mathrm{w}_j}-1}{\alpha_{\mathrm{w}_j}}\sum_{s=0}^{\mathrm{u}_j-1}\binom{\mathrm{u}_j-1}{s}(-1)^s e^{\frac{M-\mathrm{u}_j+1+s}{\rho\alpha_{\mathrm{w}_j}}}$$
$$\times E_i\left(-\frac{K-\mathrm{u}_j+1+s}{\rho\alpha_{\mathrm{w}_j}}\right)+\sum_{q=2}^{\infty}\binom{\frac{4}{K}\beta_{\mathrm{u}_j}}{q}\left(\frac{\alpha_{\mathrm{w}_j}-1}{\alpha_{\mathrm{w}_j}}\right)^q\sum_{s=0}^{\mathrm{u}_j-1}\binom{\mathrm{u}_j-1}{s}(-1)^s\Bigg(\frac{\sum_{i=1}^{q-1}\frac{(i-1)!}{(\alpha_{\mathrm{w}_j})^{-i}}\left(-\frac{K-\mathrm{u}_j+1+s}{\rho}\right)^{q-i-1}}{(q-1)!}$$
$$-\frac{\left(-\frac{K-\mathrm{u}_j+1+s}{\rho}\right)^{q-1}}{(q-1)!}\;e^{\frac{K-\mathrm{u}_j+1+s}{\rho\alpha_{\mathrm{w}_j}}}E_i\left(-\frac{K-\mathrm{u}_j+1+s}{\rho\alpha_{\mathrm{w}_j}}\right)\Bigg)\Bigg)\Bigg), \tag{7a}$$

$$E_{\mathrm{c}}^{\mathrm{w}_j}=-\frac{1}{\theta_{\mathrm{w}_j}T_{\mathrm{f}}B}\ln\left(\frac{\psi_{\mathrm{w}_j}}{\rho\alpha_{\mathrm{w}_j}}\sum_{s=0}^{\mathrm{w}_j-1}\binom{\mathrm{w}_j-1}{s}(-1)^s U\left(1,2+\frac{4}{K}\beta_{\mathrm{w}_j},\frac{K-\mathrm{w}_j+1+s}{\rho\alpha_{\mathrm{w}_j}}\right)\right), \tag{7b}$$

$$E_{\mathrm{c}}^{\mathrm{w}_j}\approx-\frac{1}{\theta_{\mathrm{w}_j}T_{\mathrm{f}}B}\ln\left(\frac{\psi_{\mathrm{w}_j}}{\rho}\left(\alpha_{\mathrm{w}_j}\right)^{\frac{4}{K}\beta_{\mathrm{w}_j}}\sum_{s=0}^{\mathrm{w}_j-1}\binom{\mathrm{w}_j-1}{s}(-1)^s\left(\frac{K-\mathrm{w}_j+1+s}{\rho}\right)^{-\frac{4}{K}\beta_{\mathrm{w}_j}-1}\Gamma\left(\frac{4}{K}\beta_{\mathrm{w}_j}+1\right)\right), \tag{7c}$$

$$\bar{E}_{\mathrm{c}}^{k}=-\frac{1}{\theta_k T_{\mathrm{f}}B}\ln\left(\frac{\psi_k}{\rho}\sum_{s=0}^{k-1}\binom{k-1}{s}(-1)^s U\left(1,2+\frac{2}{K}\beta_k,\frac{K-k+1+s}{\rho}\right)\right), \tag{7d}$$

$$\bar{E}_{\mathrm{c}}^{k}\approx-\frac{1}{\theta_k T_{\mathrm{f}}B}\ln\left(\frac{\psi_k}{\rho}\sum_{s=0}^{k-1}\binom{k-1}{s}(-1)^s\left(\frac{K-k+1+s}{\rho}\right)^{-\frac{2}{K}\beta_k-1}\Gamma\left(\frac{2}{K}\beta_k+1\right)\right). \tag{7e}$$

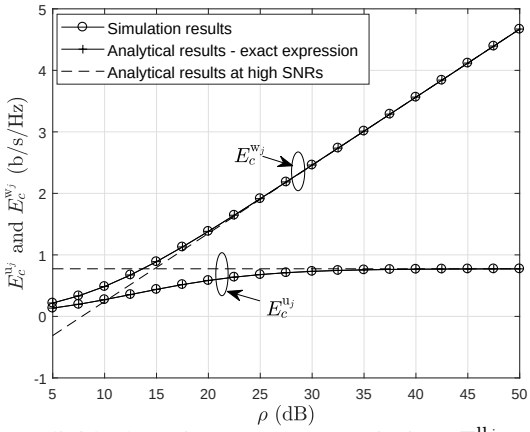


Fig. 1: Individual EC in one NOMA pair, i.e., $E_{\mathrm{c}}^{\mathrm{u}_j}$ and $E_{\mathrm{c}}^{\mathrm{w}_j}$, vs. the transmit SNR $\rho$.

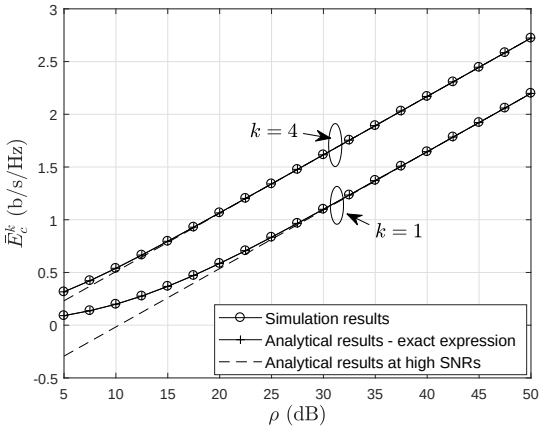


Fig. 2: Individual EC ($\bar{E}_{\mathrm{c}}^{k}$) in OMA vs. the transmit SNR $\rho$.

and $\alpha_{\mathrm{w}_j}=0.2$. Fig. 1 shows the values of $E_{\mathrm{c}}^{\mathrm{u}_j}$ and $E_{\mathrm{c}}^{\mathrm{w}_j}$, calculated from the exact analytical closed-forms using (7a)-(7b), the closed-forms at high SNRs given in Theorem 1, and also Monte carlo results. From Fig. 1, we can conclude that the exact analytical closed-forms are accurate for $E_{\mathrm{c}}^{\mathrm{u}_j}$ and $E_{\mathrm{c}}^{\mathrm{w}_j}$ in one pair. At high SNRs, the proposed approximations for $E_{\mathrm{c}}^{\mathrm{u}_j}$ and $E_{\mathrm{c}}^{\mathrm{w}_j}$ match with the calculated exact values. Furthermore, Fig. 1 also indicates that the curve of $E_{\mathrm{c}}^{\mathrm{u}_j}$ for the weaker user approaches a constant in high SNR regime, while the $E_{\mathrm{c}}^{\mathrm{w}_j}$ curve for the stronger user shows a monotonically increasing trend with $\rho$. The theoretical proof for this phenomenon is omitted in this paper. However, we refer the interested readers to [1] which provides a theoretical proof for the trends of $E_{\mathrm{c}}^{\mathrm{u}_j}$ and $E_{\mathrm{c}}^{\mathrm{w}_j}$ versus the transmit SNR $\rho$.

Correspondingly, Fig. 2 is plotted which includes the curves of $\bar{E}_{\mathrm{c}}^{k}$ for the $1^{\mathrm{st}}$ user and the $4^{\mathrm{th}}$ user in a downlink OMA network. The results are calculated using the exact analytical closed-form given in (7d), the closed-form at high SNRs given in (7e), and also the Monte carlo results. From Fig. 2, one can notice that the derived exact analytical expression is accurate for $\bar{E}_{\mathrm{c}}^{k}$ in a downlink OMA network. In high SNR regime,

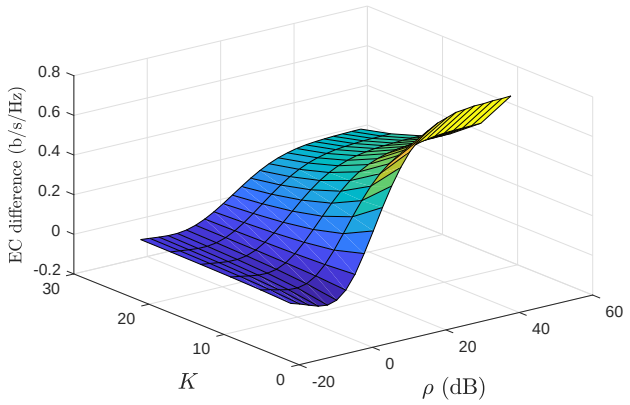


Fig. 3: $(E_c^1 + E_c^K) - (\bar{E}_c^1 + \bar{E}_c^K)$ vs. the user group size $K$ and the transmit SNR $\rho$.

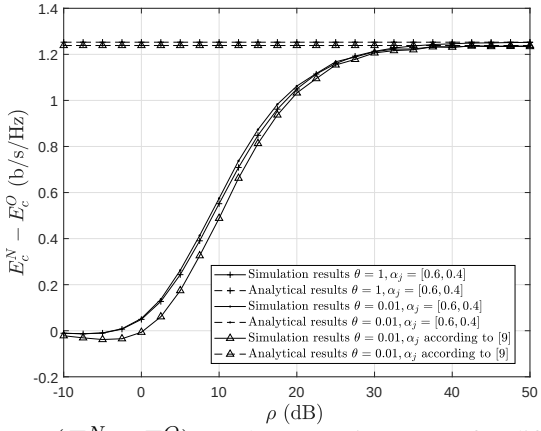


Fig. 4: $(\boldsymbol{E}_{\mathrm{c}}^N - \boldsymbol{E}_{\mathrm{c}}^O)$ vs. the transmit SNR $\rho$, for different power coefficient and $\theta$ settings.

the approximated closed-form matches with the exact values calculated with the Monte carlo method. Fig. 2 also shows that the $4^{\text{th}}$ user always achieves higher EC values. This is because that all channels have been assumed to be ordered, hence the $4^{\text{th}}$ user always has higher channel gains than the $1^{\text{st}}$ user.

It was proved in [5] that the performance gain of NOMA can be improved by pairing two users with very different channel conditions. Inspired by [5], we assume that the user with the strongest channel conditions, i.e., the user $K$, is paired with the user with the weakest channel conditions, i.e., the user 1. Then, we plot Fig. 3 to study the total EC difference for the pair $(1, K)$, i.e., $(E_c^1 + E_c^K) - (\bar{E}_c^1 + \bar{E}_c^K)$, with respect to the user group size and the transmit SNR. From Fig. 3, it is evident that although the EC difference for the pair $(1, K)$ reduces with the group size $K$, it remains positive even for large group sizes. This indicates that when the user group size becomes larger, although the performance of NOMA and OMA reduces, the spectrum sharing gain by applying NOMA achieves higher total EC at high SNRs than OMA.

Assume that all 6 NOMA users are paired as follows: $\boldsymbol{\phi}_1 = (1, 5)$, $\boldsymbol{\phi}_2 = (2, 4)$ and $\boldsymbol{\phi}_3 = (3, 6)$. All NOMA pairs have the same power coefficient settings, i.e., $\alpha_j$ is the same for $j \in \{1, 2, 3\}$. Fig. 4 illustrates the curves of $\boldsymbol{E}_{\mathrm{c}}^N - \boldsymbol{E}_{\mathrm{c}}^O$ versus $\rho$, for different delay requirement scenarios and power coefficient settings within one pair. The solid lines are calculated with Monte carlo results and the dashed lines are plotted using the derived analytical closed-forms at high SNRs. Firstly, Fig. 4 shows that when all users have a slightly stringent delay requirement, i.e., when delay exponent $\theta$ changes from 0.01 to 1, the values of $\boldsymbol{E}_{\mathrm{c}}^N - \boldsymbol{E}_{\mathrm{c}}^O$ do not vary very much. But since $\theta$ indicates the exponential decay rate of the delay outage probability, with a slight increase of $\theta$, the delay outage probability decreases in an exponential way. Secondly, two different power settings are included in Fig. 4 for the two users whithin one pair, which are a fixed power allocation $\boldsymbol{\alpha}_j = (0.6, 0.4)$, $\forall j \in \{1, 2, 3\}$, and a varied power coefficient setting according to [9]. Fig. 4 shows that the two employed power settings achieve different $\boldsymbol{E}_{\mathrm{c}}^N - \boldsymbol{E}_{\mathrm{c}}^O$ values, but observe the same trend of $\boldsymbol{E}_{\mathrm{c}}^N - \boldsymbol{E}_{\mathrm{c}}^O$ versus $\rho$. Finally, Fig. 4 also indicates that the $\boldsymbol{E}_{\mathrm{c}}^N - \boldsymbol{E}_{\mathrm{c}}^O$ values at high SNRs converge to a constant, which confirms the theoretical analysis in Lemma 1.

## V. Conclusions

In order to investigate the performance of different multiple access protocols for delay-sensitive applications, we focused on the downlink transmission which utilizes either user-paired NOMA principle or conventional OMA technique. It was proved that the performance gain of NOMA over OMA on the total delay-constrained rate converges at very high SNRs. This means that the further increase in SNR will lead to diminishing returns in the total delay-constrained rate. Further, for a special user pair with the two users having the most distinctive channel conditions, it was shown that at high SNRs, NOMA always achieves higher total EC even for a very large group size.

## VI. Appendix

Proof for Theorem 1: In high SNR regime, $E_{\mathrm{c}}^{\mathrm{u}_j}$ for the weaker user in the $j^{\text{th}}$ NOMA pair, given in (6a), can be approximated as follows:

$$\lim_{\rho\to\infty} E_{\mathrm{c}}^{\mathrm{u}_j} = -\frac{1}{\theta_{\mathrm{u}_j} T_{\mathrm{f}} B} \ln\left(\frac{1}{\alpha_{\mathrm{w}_j}}\right)^{\frac{4}{K}\beta_{\mathrm{u}_j}} = \frac{2}{K}\log_2\frac{1}{\alpha_{\mathrm{w}_j}}. \quad (11)$$

At high SNRs, $E_{\mathrm{c}}^{\mathrm{w}_j}$ for the stronger user in the $j^{\text{th}}$ NOMA pair, given in (6b), can be approximated and written as

$$E_{\mathrm{c}}^{\mathrm{w}_j} \approx -\frac{1}{\theta_{\mathrm{w}_j} T_{\mathrm{f}} B} \ln\left(\mathbb{E}\left[\left(\rho\alpha_{\mathrm{w}_j}|h_{\mathrm{w}_j}|^2\right)^{\frac{4}{K}\beta_{\mathrm{w}_j}}\right]\right). \quad (12)$$

By setting $x_{\mathrm{w}_j} = \rho|h_{\mathrm{w}_j}|^2$ and inserting $f(x_{\mathrm{w}_j}) = \frac{1}{\rho}e^{-\frac{x_{\mathrm{w}_j}}{\rho}}$, $F(x_{\mathrm{w}_j}) = 1 - e^{-\frac{x_{\mathrm{w}_j}}{\rho}}$ into (5), (12) can be expanded as

$$E_{\mathrm{c}}^{\mathrm{w}_j} \approx -\frac{1}{\theta_{\mathrm{w}_j} T_{\mathrm{f}} B} \ln\left(\frac{\psi_{\mathrm{w}_j}}{\rho}\left(\alpha_{\mathrm{w}_j}\right)^{\frac{4}{K}\beta_{\mathrm{w}_j}} \int_0^\infty (x_{\mathrm{w}_j})^{\frac{4}{K}\beta_{\mathrm{w}_j}} \times e^{-\frac{(K-\mathrm{w}_j+1)x_{\mathrm{w}_j}}{\rho}} \left(1 - e^{-\frac{x_{\mathrm{w}_j}}{\rho}}\right)^{\mathrm{w}_j-1} dx_{\mathrm{w}_j}\right) \quad (13)$$

$$= -\frac{1}{\theta_{\mathrm{w}_j} T_{\mathrm{f}} B} \ln \left( \frac{\psi_{\mathrm{w}_j}}{\rho} \left(\alpha_{\mathrm{w}_j}\right) \frac{4}{K}^{\beta_{\mathrm{w}_j}} \sum_{s=0}^{\mathrm{w}_j - 1} \binom{\mathrm{w}_j - 1}{s} (-1)^s \right.$$
$$\left. \times \int_0^{\infty} (x_{\mathrm{w}_j}) \frac{4}{K}^{\beta_{\mathrm{w}_j}} e^{-\frac{(K - \mathrm{w}_j + 1 + s) x_{\mathrm{w}_j}}{\rho}} dx_{\mathrm{w}_j} \right) \quad (14)$$

$$= -\frac{1}{\theta_{\mathrm{w}_j} T_{\mathrm{f}} B} \ln \left( \frac{\psi_{\mathrm{w}_j}}{\rho} \left(\alpha_{\mathrm{w}_j}\right) \frac{4}{K}^{\beta_{\mathrm{w}_j}} \sum_{s=0}^{\mathrm{w}_j - 1} \binom{\mathrm{w}_j - 1}{s} (-1)^s \right.$$
$$\left. \times \left( \frac{K - \mathrm{w}_j + 1 + s}{\rho} \right)^{-\frac{4}{K}^{\beta_{\mathrm{w}_j} - 1}} \Gamma\left( \frac{4}{K} \beta_{\mathrm{w}_j} + 1 \right) \right). \quad (15)$$

From (13) to (14), it is obtained by replacing $\left(1 - e^{-\frac{x_{\mathrm{w}_j}}{\rho}}\right)^{\mathrm{w}_j - 1}$ with its binomial expansion $\sum_{s=0}^{\mathrm{w}_j - 1} \binom{\mathrm{w}_j - 1}{s} (-1)^s e^{-\frac{x_{\mathrm{w}_j}}{\rho} s}$. Then, from (3.382.4) in [18], we have that

$$\int_0^{\infty} (x + \beta)^v e^{-ux} dx = u^{-v-1} e^{\beta u} \Gamma(v + 1, \beta u)$$
$$\text{for } |\arg \beta| < \pi, \mathrm{Re}\ u > 0. \quad (16)$$

By applying (16) to (14) and converting the incomplete gamma function $\Gamma\left(\frac{4}{K}\beta_{\mathrm{w}_j} + 1, 0\right)$ to $\Gamma\left(\frac{4}{K}\beta_{\mathrm{w}_j} + 1\right)$, (15) can be finally derived.

Proof for Theorem 2: At high SNRs, $\bar{E}_{\mathrm{c}}^k$ for the $k^{\mathrm{th}}$ user in a conventional OMA network, given in (8), can be approximated and written as

$$\bar{E}_{\mathrm{c}}^k \approx -\frac{1}{\theta_k T_{\mathrm{f}} B} \ln \left( \mathbb{E} \left[ \left(\rho |h_k|^2\right) \frac{2}{K}^{\beta_k} \right] \right). \quad (17)$$

By setting $y_k = \rho |h_k|^2$ and inserting the PDF $f_{(k)}(y_k)$, (17) can be expanded as

$$\bar{E}_{\mathrm{c}}^k \approx -\frac{1}{\theta_k T_{\mathrm{f}} B} \ln \left( \frac{\psi_k}{\rho} \int_0^{\infty} (y_k) \frac{2}{K}^{\beta_k} e^{-\frac{(K - k + 1) y_k}{\rho}} \right.$$
$$\left. \times \left(1 - e^{-\frac{y_k}{\rho}}\right)^{k-1} dy_k \right) \quad (18)$$

$$= -\frac{1}{\theta_k T_{\mathrm{f}} B} \ln \left( \frac{\psi_k}{\rho} \sum_{s=0}^{k-1} \binom{k-1}{s} (-1)^s \right.$$
$$\left. \times \left( \frac{K - k + 1 + s}{\rho} \right)^{-\frac{2}{K}^{\beta_k - 1}} \Gamma\left( \frac{2}{K} \beta_k + 1 \right) \right). \quad (19)$$

From (18) to (19), it is achieved by first substituting $\left(1 - e^{-\frac{y_k}{\rho}}\right)^{k-1}$ with its binomial expansion and then applying (16).

## VII. Acknowledgement

This work was partly funded by grant N62909-17-1-2114 from the US Office of Naval Research Global, the Royal Society project IEC170324, the EPSRC project EP/K011693/1, and EU FP7 CROWN project PIRSES-GA-2013-610524. In addition, the authors would also like to acknowledge the support of the University of Surrey 5GIC (http://www.surrey.ac.uk/5gic) members for this work.

## References

[1] W. Yu, L. Musavian, and Q. Ni, "Link-layer capacity of NOMA under statistical delay QoS guarantees," *IEEE Trans. Commun.*, in press.

[2] C. She, C. Yang, and T. Q. S. Quek, "Cross-layer optimization for ultra-reliable and low-latency radio access networks," *IEEE Trans. Wirel. Commun.*, vol. 17, no. 1, pp. 127–141, Jan. 2018.

[3] W. Yu, L. Musavian, and Q. Ni, "Tradeoff analysis and joint optimization of link-layer energy efficiency and effective capacity toward green communications," *IEEE Trans. Wirel. Commun.*, vol. 15, no. 5, pp. 3339–3353, Jan. 2016.

[4] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo, "Nonorthogonal multiple access for 5G and beyond," *Proc. IEEE*, vol. 105, no. 12, pp. 2347–2381, Dec. 2017.

[5] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.

[6] P. Xu and K. Cumanan, "Optimal power allocation scheme for non-orthogonal multiple access with $\alpha$-fairness," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2357–2369, Oct. 2017.

[7] L. Lv, J. Chen, Q. Ni, and Z. Ding, "Design of cooperative non-orthogonal multicast cognitive multiple access for 5G systems: User scheduling and performance analysis," *IEEE Trans. Commun.*, vol. 65, no. 6, pp. 2641–2656, Jun. 2017.

[8] Y. Liu, Z. Qin, M. Elkashlan, Y. Gao, and L. Hanzo, "Enhancing the physical layer security of non-orthogonal multiple access in large-scale networks," *IEEE Trans. Wirel. Commun.*, vol. 16, no. 3, pp. 1656–1672, Jan. 2017.

[9] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.

[10] M. Gursoy, D. Qiao, and S. Velipasalar, "Analysis of energy efficiency in fading channels under QoS constraints," *IEEE Trans. Wirel. Commun.*, vol. 8, no. 8, pp. 4252–4263, Aug. 2009.

[11] J. Choi, "Effective capacity of NOMA and a suboptimal power control policy with delay QoS," *IEEE Trans. Commun.*, vol. 65, no. 4, pp. 1849–1858, Jan. 2017.

[12] G. Liu, Z. Ma, X. Chen, Z. Ding, F. R. Yu, and P. Fan, "Cross-layer power allocation in nonorthogonal multiple access systems for statistical qos provisioning," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 11 388–11 393, Dec. 2017.

[13] C. S. Chang, "Stability, queue length and delay, part ii: Stochastic queueing networks," in *Proc. IEEE Conf. Decision Contr.*, vol. 1, Tucson, Arizona, USA, Dec. 1992, pp. 1005–1010.

[14] J. T. Lewis and R. Russell, *An Introduction to Large Deviations for Teletraffic Engineers*, 1997.

[15] J. A. Bucklew, *Introduction to Rare Event Simulation*. Springer-Verlag New York Inc., 2004.

[16] H. A. David and H. N. Nagaraja, *Order Statistics*. John Wiley, New York, 3rd ed., 2003.

[17] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*. New York: Dover, 1965.

[18] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*. New York: Academic Press, 6th ed., 2000.