Minimizing the Transaction Time Difference for NOMA-Based Mobile Edge Computing

Anam Yasir Kiani, Syed Ali Hassan, Binbin Su, Haris Pervaiz and Qiang Ni

***Abstract*—Non orthogonal multiple access (NOMA) and mobile edge computing (MEC) are evolving as key enablers for fifth generation (5G) networks as this combination can provide high spectral efficiency, improved quality-of-service (QoS), and lower latency. This letter aims to minimize the transaction time difference of two NOMA paired users offloading data to MEC servers by optimizing their transmission powers and computational resources of severs using a successive convex approximation method. The equalization of transaction time for paired users reduces the wastage of both frequency and computational resources, and improves effective throughput of the system to 19% on average.**

***Index Terms*—Non-orthogonal multiple access (NOMA), Mobile Edge Computing (MEC), 5G communications**

I. Introduction

Some of the important concerns in today's wireless networks are limited resources and media arbitration. The medium used for the transmissions is being shared by millions of devices with heavy traffic, which is expected to increase by 1000 folds in the next decade. This would result in services requiring high connectivity, reliability, ultra-low latency, improved fairness and high throughput, etc. Non-orthogonal multiple access (NOMA) has been introduced to muddle through the demands of the epoch. One of main purposes of NOMA is to serve multiple users by utilizing the same resource block. NOMA provides a balanced trade-off between the system throughput and user fairness [1] and is being envisioned as a key technology in 5G networks. 5G enabled devices are also expected to have latency constraint and have computationally complex applications running on them. For such applications, limited power and computational capacity of mobile devices pose a problem, which can be solved by using mobile edge computing (MEC) [2]. MEC offloads computationally intensive data to base stations (BSs) and access points (APs) that are equipped with powerful servers. Servers being available at the edges result in reduction of delay and improvement of computational efficiency [3]. The advantages of both techniques (i.e., NOMA and MEC) have drawn considerable attention of the researchers recently. In NOMA-MEC, paired users offload their data to MEC servers by using the underlying NOMA principle.

A. Y. Kiani and S. A. Hassan are with the School of Electrical Engineering & Computer Science (SEECS), National University of Sciences & Technology (NUST), Islamabad, Pakistan. e-mails: {akiani.msee17seecs, ali.hassan}@seecs.edu.pk. B. Su, H. Pervaiz and Q. Ni are with School of Computing & Communications, Lancaster University, UK. e-mails: {b.su, h.b.pervaiz, q.ni}@lancaster.ac.uk. This work was supported in part by the EPSRC Global Challenges Research Fund (GCRF) DARE project grant no. EP/P028764/1.

A lot of work is being done in this context. For instance, [4] formulated delay minimization for NOMA-MEC data offloading as a form of fractional programming. Pure NOMA is also compared with hybrid NOMA and orthogonal multiple access (OMA) for data offloading purpose. In [5], energy consumption of MEC users utilizing uplink NOMA is reduced by optimizing user clustering, power, frequency and computational resource allocation. [6] proved that the total energy minimization is a convex problem and the authors solved it by an iterative algorithm. [7] reduced the total system energy by optimizing allocated power, transmission time and offloaded task portions. In [8], energy consumption is reduced by jointly optimizing power and time allocation by formulating the problem to a form of geometric programming. [9] studied energy harvesting for full duplex NOMA-MEC, where total energy consumption is minimized by efficient power allocation, time scheduling and computing resources allocation.

The time taken to process data (offloaded to MEC servers) is not equal for each NOMA paired user because it is dependent upon the amount of offloaded data and the channel conditions of the paired users, etc. This inequality leads to under-utilization of resources and reduced spectral efficiency. In this letter, we propose a scheme to optimize the transaction time of paired users and to reduce the transaction time difference between them to improve spectral efficiency and to conserve both frequency and computational resources. Transaction time is the sum of transmission time and computational time. The difference between transaction times is reduced by equalizing transmission time and computational time separately, which is achieved by optimizing a) power allocation and b) computational resource allocation, respectively. When the transaction time of paired users becomes closer, the difference between the transaction times reduces and hence the wasteful resources.

In the sequel, we describe the NOMA-MEC model followed by optimization of transaction time difference. The results and conclusions are presented towards the end.

II. System Model

A single cell is considered with $2N$ number of users, which are served by a single BS. The BS is equipped with MEC server having C_T number of cores each with a computational capability of f cycles/sec and total system bandwidth is B_T. Hybrid NOMA technique is used to pair the users into N NOMA clusters, where each cluster has two users. A single cluster with users u_1 and u_2 is considered for study. The user u_1 is located at a distance dist_1 from BS whereas u_2 is located

at a distance dist_2 from BS, such that $\text{dist}_1 < \text{dist}_2$. Without the loss in generality, we assume that u_1 is a strong user with allocated power p_1 and effective channel gain h_1, however u_2 is a weak user with allocated power p_2 and effective channel gain h_2, such that $p_1 h_1 > p_2 h_2$. Let $p_{1,\text{max}}$ and $p_{2,\text{max}}$ are the maximum transmission powers that can be allocated to u_1 and u_2, respectively. We assume that d_1 bits are offloaded by u_1 and d_2 bits are offloaded by u_2 to the MEC server. Complete offloading scheme is considered, where no local computation is being performed. Each bit offloaded by u_1 requires c_1 cycles and that of u_2 requires c_2 cycles for computation at MEC server. The computational complexity of offloaded data is dependent upon offloaded data type (i.e, video data requires more CPU cycles as compared to text data). The total system bandwidth is divided into N number of frequency resource blocks. A single frequency resource block with bandwidth $B_w = B_T/N$ is allocated to a NOMA cluster and shared by paired users, similarly the cores at MEC servers are divided into N number of computational resource blocks and are allocated to NOMA clusters. The allocated computational resources of a cluster (i.e., $C_t = C_T/N$) are divided among the paired users, depending upon the complexity and amount of data being offloaded by them. Let u_1 and u_2 are allocated with n_1 and n_2 cores, respectively.

The transaction time of the i^{th} user is $T_i = T_{tx_i} + T_{c_i}, \quad i \in \{1, 2\}$, where T_{tx_i} is the transmission time and T_{c_i} is the computational time of the i^{th} user, respectively. The transmission time for the i^{th} user is $T_{tx_i} = \frac{d_i}{R_i}$, where R_i is the data rate of the i^{th} user. In this work, the 2-user uplink NOMA cluster is considered in which u_1 and u_2 experience channel gains of $\widetilde{h_1}$ and $\widetilde{h_2}$ such that the user u_1 signal will be decoded first at BS. The achievable data rate of user u_1 will include the interference from the user u_2 whereas the achievable data rate of user u_2 will include noise only. The data rates are dependent upon the effective channel gains (i.e., h_1, h_2) and the allocated powers (i.e., p_1, p_2), such that [10]

$$R_1 = B_w \log_2 \left(1 + \frac{p_1 h_1}{p_2 h_2 + \sigma^2}\right), \tag{1}$$

$$R_2 = B_w \log_2 \left(1 + \frac{p_2 h_2}{\sigma^2}\right), \tag{2}$$

where $\sigma^2 = B_w \times \overline{\sigma}^2$, $\overline{\sigma}^2$ is the power spectral density of noise and the effective channel gain for i^{th} user is $h_i = \frac{\widetilde{h_i}}{dist_i^{\rho}}$, where $\widetilde{h_i}$ is the exponential channel gain (corresponding to Rayleigh fading) of i^{th} user and ρ is the path loss exponent. The amount of data offloaded and the effective channel gains are associated with the paired users, however, the powers are optimized to reduce their transaction time difference. Similarly, the computational time for the i^{th} user is given by $T_{c_i} = \frac{d_i c_i}{n_i f}$, where n_i is the number of cores allocated to the user i and f is the computational capacity of each MEC core. For a given paired users, d_i, c_i and f are fixed. The number of computational resources allocated to the i^{th} user is optimized to balance the load across the cores in order to reduce the

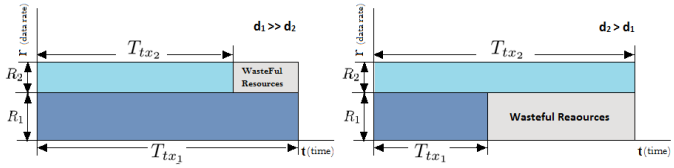

Fig. 1: Unequal Transmission Time and Wasteful Resources

difference between transaction time. By manipulating, it is inferred that T_1 is equal to T_2 if

$$\begin{aligned}
\frac{d_1}{R_1} + \frac{d_1 c_1}{n_1 f} &= \frac{d_2}{R_2} + \frac{d_2 c_2}{n_2 f} \\
\frac{n_1 f d_1 + d_1 c_1 R_1}{n_1 f R_1} &= \frac{d_2 n_2 f + d_2 c_2 R_2}{n_2 f R_2} \\
\frac{d_1 \left(n_1 f + c_1 R_1\right)}{n_1 R_1} &= \frac{d_2 \left(n_2 f + c_2 R_2\right)}{n_2 R_2} \\
\frac{d_1}{d_2} &= \left(\frac{R_1}{R_2}\right) \left(\frac{n_1 n_2 f + n_1 c_2 R_2}{n_1 n_2 f + n_2 c_1 R_1}\right)
\end{aligned} \tag{3}$$

From (3), we can divide the original formulated problem into two independent sub-problems and reformulate it as T_1 is equal to T_2, if T_{tx_1} is equal to T_{tx_2} as well as T_{c_1} is equal to T_{c_2}. It is evident from Fig. 1 that unequal transmission time results in wastage of allocated frequency resources. It can be seen that for $\delta_{tx} = |T_{tx_1} - T_{tx_2}|$ amount of time, the resources are under-utilized, i.e., a new NOMA signal cannot be initiated. As this difference increases, the spectral efficiency of the network decreases. The transmission time, T_{tx}, is equal for both the users, if $\frac{d_1}{d_2} = \frac{R_1}{R_2}$, where R_1 and R_2 are the data rates of both users u_1 and u_2, respectively. Similarly, the disparity in amount and computational complexity of data offloaded by paired users (i.e., allocated with equal number of cores) results in wastage of allocated computational resources. The computational time difference $\delta_{tc} = |T_{tc1} - T_{tc2}|$ for both the users is zero, if $\frac{d_1 c_1}{n_1 f} = \frac{d_2 c_2}{n_2 f}$, which can be written in simplified form as $\frac{d_1}{d_2} = \frac{n_1 c_2}{n_2 c_1}$.

III. Problem Formulation

The original problem described in the previous section is given by

$$(\mathbf{P}) \quad \min \quad \lambda, \tag{4a}$$

$$s.t. \quad \left(\frac{d_1}{R_1} + \frac{d_1 c_1}{n_1}\right) - \left(\frac{d_2}{R_2} + \frac{d_2 c_2}{n_2}\right) \leq \lambda, \tag{4b}$$

$$\left(\frac{d_2}{R_2} + \frac{d_2 c_2}{n_2}\right) - \left(\frac{d_1}{R_1} + \frac{d_1 c_1}{n_1}\right) \leq \lambda, \tag{4c}$$

$$\lambda \geq 0 \tag{4d}$$

$$0 \leq p_i \leq p_{i,\text{max}} \quad i \in \{1, 2\} \tag{4e}$$

$$0 \leq n_i \leq C_t \quad i \in \{1, 2\} \tag{4f}$$

$$n_1 + n_2 \leq C_t \tag{4g}$$

where the problem (P) is subjected to constraints (4e), (4f) and (4g) i.e., the power allocated to the individual user is

positive and less than respective maximum, the number of cores allocated to the individual user is positive and less than total number of allocated cores, moreover, sum of cores allocated to both the users is less than or equal to the total number of allocated cores. As can be seen from the formation of problem (P), it can be decomposed into two independent optimization sub-problems. The results for original optimization problem and sub-problems are equivalent. The objective of the first optimization problem is to minimize the transmission time difference of given paired users with known d_i's, by optimizing the power allocation. From equations (1) and (2), we have

$$R_1 + R_2 = B_w \log_2 \left(1 + \frac{p_1 h_1 + p_2 h_2}{\sigma^2}\right) \tag{5a}$$

$$\begin{aligned} R_1 &\leq B_w \log_2 \left(1 + \frac{p_1 h_1}{p_2 h_2 + \sigma^2}\right) \\ &= B_w \log_2 \left(1 + \frac{p_1 h_1 + p_2 h_2}{\sigma^2}\right) - R_2, \end{aligned} \tag{5b}$$

$$R_2 \leq B_w \log_2 \left(1 + \frac{p_2 h_2}{\sigma^2}\right), \tag{5c}$$

For the objective with power allocation, we introduce a new variable μ and hence the sub-problem of minimizing the transmission time difference of paired users can be reformulated as

$$(\mathbf{P1}) \qquad \min \quad \mu, \tag{6a}$$

$$s.t. \qquad \frac{d_1}{R_1} - \frac{d_2}{R_2} \leq \mu, \tag{6b}$$

$$\frac{d_2}{R_2} - \frac{d_1}{R_1} \leq \mu, \tag{6c}$$

$$\mu \geq 0 \tag{6d}$$

$$0 \leq p_i \leq p_{i,\max} \quad i \in \{1, 2\} \tag{6e}$$

where the objective function (6a) is subjected to data rate (5b, 5c) and power (6e) constraints. By manipulating (6b), we get

$$\mu R_1 R_2 \geq \mu \alpha_1 \geq \alpha_2^2 \geq d_1 R_2 - d_2 R_1, \tag{7}$$

where α_1 and α_2 are real valued variables, having values such that inequality holds. The equation (7) is equivalent to

$$R_1 R_2 \geq \alpha_1, \tag{8a}$$

$$\begin{bmatrix} \mu & \alpha_2 \\ \alpha_2 & \alpha_1 \end{bmatrix} \succeq 0, \tag{8b}$$

$$\alpha_2^2 \geq d_1 R_2 - d_2 R_1, \tag{8c}$$

where (8b) is a convex linear matrix inequality (LMI), and (8c) is non-convex. The non-convex parts in left side of (8c) can be approximated using the Taylor series expansion to get the approximated lower bound. By applying the first Order Taylor Approximation, the left side of (8c) can be approximated as

$$\alpha_2{}^2 \geq \left(\alpha_2^{(j)}\right)^2 + 2\alpha_2^{(j)} \left(\alpha_2 - \alpha_2^{(j)}\right)$$

$$\alpha_2{}^2 \geq \left(\alpha_2^{(j)}\right)^2 + 2\alpha_2^{(j)} \alpha_2 - 2\left(\alpha_2^{(j)}\right)^2$$

$$\alpha_2{}^2 \geq 2\alpha_2^{(j)} \alpha_2 - \left(\alpha_2^{(j)}\right)^2 \tag{9}$$

The right side of Eq. (9) is the first order approximation around the point $\left(\alpha_2^{(j)}\right)$. By substituting Eq. (9) into left side of (8c), the (8c) can be rewritten as follows:

$$2\alpha_2^{(j)} \alpha_2 - (\alpha_2^{(j)})^2 \geq d_1 R_2 - d_2 R_1, \tag{10}$$

where j shows the number of iteration, $\alpha_2^{(j)}$ denotes the value of α_2 during the $j-th$ iteration. The equation (8a) is rewritten as

$$R_1 R_2 \geq \beta^2, \tag{11}$$

where $\beta^2 \geq \alpha_1$. The problem (**P1**) defined in Eq. (6a) subject to the constraints defined in Eq. (8b), Eq. (10) and Eq. (11) is a convex optimization problem and can be efficiently solved using standard convex optimization tool such as CVX [11]. It will provide a lower bound approximation solution [12], [13] of (**P1**) due to the first order Taylor approximation in Eq. (10). Similarly, the objective of the second optimization problem is to minimize the computational time difference of given paired users with known d_i's and c_i's by optimizing the core allocation. By introducing a new variable ζ, the sub-problem of the computational resource allocation can be transformed as

$$(\mathbf{P2}) \qquad \min \quad \zeta, \tag{12a}$$

$$s.t. \qquad \frac{d_1 c_1}{n_1} - \frac{d_2 c_2}{n_2} \leq \zeta, \tag{12b}$$

$$\frac{d_2 c_2}{n_2} - \frac{d_1 c_1}{n_1} \leq \zeta, \tag{12c}$$

$$\zeta \geq 0, \tag{12d}$$

$$0 < n_i < C_t, \tag{12e}$$

$$n_1 + n_2 \leq C_t \tag{12f}$$

where the objective function (12a) is subject to constraints (12e), the number of cores allocated to individual user is greater than zero and less than total cores allocated to the cluster and (12f), the sum of cores allocated to both the users is less than or equal to total cores allocated to the cluster. The number of cores allocated to individual user must be greater than zero to ensure the minimum requirement of the user. The integer constraint is relaxed for n_i. By manipulating (12b), we get

$$\zeta n_1 n_2 \geq \zeta \gamma_1 \geq \gamma_2^2 \geq (d_1 c_1) n_2 - (d_2 c_2) n_1, \tag{13}$$

where γ_1 and γ_2 are variables with real values. The equation (13) implies

$$n_1 n_2 \geq \gamma_1, \tag{14a}$$

$$\begin{bmatrix} \zeta & \gamma_2 \\ \gamma_2 & \gamma_1 \end{bmatrix} \succeq 0, \tag{14b}$$

$$\gamma_2^2 \geq (d_1 c_1) n_2 - (d_2 c_2) n_1, \tag{14c}$$

where (14b) is a convex LMI. The left side of (14c) can be approximated using the Taylor series expansion to get the

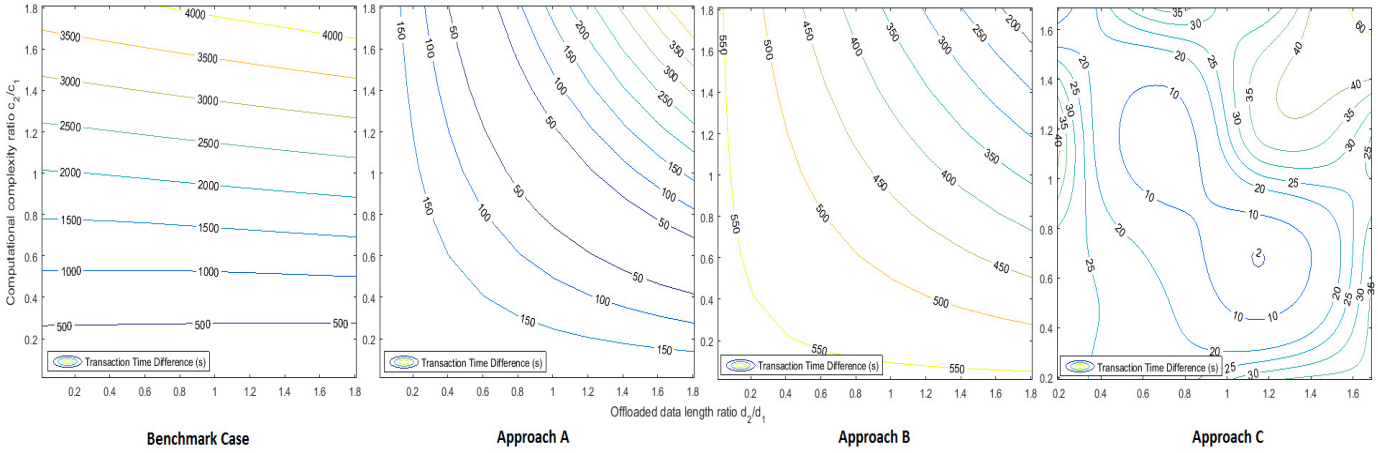

Fig. 2: Transaction Time difference for Benchmark (or No Optimization) Case, Approaches A, B and C

approximated lower bound. By applying the first Order Taylor Approximation, the left side of (14c) can be approximated as

$$\gamma_2{}^2 \geq 2\gamma_2^{(j)}\gamma_2 - \left(\gamma_2^{(j)}\right)^2 \tag{15}$$

The right side of Eq. (15) is the first order approximation around the point $\left(\gamma_2^{(j)}\right)$. By substituting Eq. (15) into left side of (14c), the (14c) can be rewritten as follows:

$$2\gamma_2^{(j)}\gamma_2 - \left(\gamma_2^{(j)}\right)^2 \geq (d_1c_1)n_2 - (d_2c_2)n_1, \tag{16}$$

where γ_i's are updated in each iteration and j shows the number of iteration. From (14a), we have

$$n_1 n_2 \geq \eta^2, \tag{17}$$

where $\eta^2 \geq \gamma_1$. The problem (**P2**) defined in Eq. (12a) subject to the constraints defined in Eq. (14b), Eq. (16) and Eq. (17) is a convex optimization problem and can be efficiently solved using standard convex optimization tool such as CVX [11].

For a given pair of users, we obtain optimal values of power and number of cores once the optimization is performed. These parameters result in minimization of transaction time difference, which is illustrated in next section.

IV. Performance Evaluation

The maximum power for u_1, $p_{1,\max}$, is 2W and of u_2, $p_{2,\max}$, is 4W. Initially p_1 is 1W and p_2 is 2W. The $dist_1$ and $dist_2$ are 200 m and 600 m, respectively. The cluster bandwidth is 200 kHz and the path loss exponent is 3.8. The transaction time difference is considered for three different approaches namely: Power Optimization with Equal Core Allocation (A), Power Optimization with Random Core Allocation (B) and proposed Power Optimization with Optimal Core Allocation (C). The power and core optimization is achieved by successive convex approximation as discussed in the Section III. In equal core allocation, the cores are equally divided between the paired users, i.e., $n_1 = n_2$. In random core allocation, the cores are randomly divided between the paired users $n_1 = \kappa C_t$ and $n_2 = (1-\kappa)C_t$, where, κ is from uniform random distribution varying from 0 to 1. The ratio of offloaded data amount, i.e., d_2/d_1 and complexity, i.e., c_2/c_1 is varied to study their impact on the transaction time.

Fig. 2 depicts the transaction time difference without any power optimization and equal number of core allocation, i.e., benchmark case, approach A, approach B and approach C in contour plots from left to right. It can be observed in all plots that for a fixed value of d_2/d_1, different values of c_2/c_1 result in different transaction time differences. The larger the transaction time difference, the more the under-utilized resources. It can be observed that the transaction time difference for second plot (i.e., approach A) is overall lesser than the previous. For the same ratios of d_2/d_1 and c_2/c_1, the transaction time difference is reduced by optimizing only the power allocations. The transaction time difference for Approach B (i.e., third plot) is lesser than the transaction time difference for first plot, i.e., benchmark case. However, this difference is comparable with Approach A, as the only difference is in the core allocation. The transaction time difference for proposed scheme (i.e., Approach C, fourth plot in Fig. 2), where both the power and cores are optimized, is minimum. It is also clear that the paired users have optimal values of d_2/d_1 and c_2/c_1 for which the transaction time difference is minimum. For instance in fourth plot Fig. 2, when $d_2/d_1 = 1.2$, the transaction time difference is 2 seconds for c_2/c_1 of 0.7. As d_2/d_1 is increased to 1.7, the transaction time gap jumps to 35 seconds for the same ratio of c_2/c_1. Similarly, when d_2/d_1 is decreased to 0.7, the transaction time difference increases to 20 seconds.

To validate the proposed solution, the approaches A, B and C are also solved heuristically by searching over the whole solution space labelled as "Simulation" and compared with the results obtained for the approaches A, B and C using successive convex optimization method labelled as "Optimization" in Fig.3. The maximum number of iterations for the Optimization results for the Approaches A, B and C is set to 100. In Fig.3 when d_2/d_1 is 0.4, the transaction time difference without optimization is 822 seconds, approximately 402 seconds for both heuristic and optimized solutions of Approach A and 467 seconds for Approach B. The transaction time difference for Approach C goes to 196.1 and 61.58 seconds for optimized

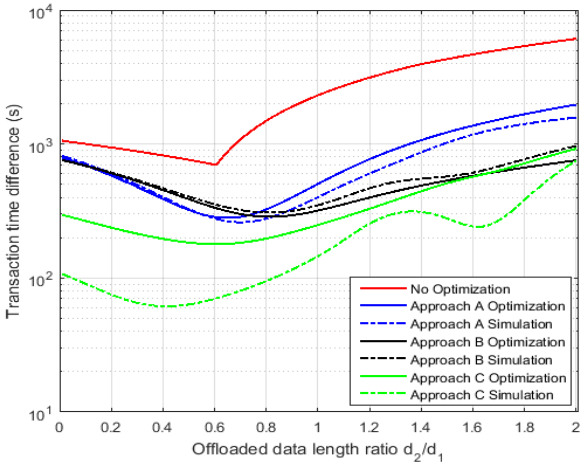

Fig. 3: Comparison of Simulation and Optimization Results

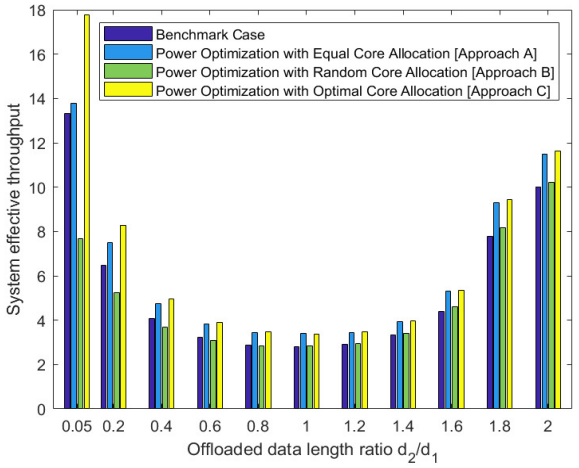

Fig. 4: Effective System Throughput

and heuristic solutions,respectively. The difference between the simulation and optimization results is due to use of Taylor series expansion for approximation.

We now illustrate the effect of reducing the transaction time on the *effective throughput* of the system. The effective throughput of the system is given by

$$\Phi_{eff} = \frac{\sum_{i=1}^{2} R_i}{\max(T_1, T_2)}, \tag{18}$$

where the numerator is the sum of achieved data rates by the paired users while the denominator is the maximum of transaction times of the paired users. As both the users are paired, therefore, the resources allocated to them are free only when both of them complete their transactions, hence the denominator is characterized by the max(.) operator. A decrease in effective transaction time increases the system's effective throughput. It is clear from Fig. 4 that for a fixed value of c_2/c_1 and a range of d_2/d_1, the system effective throughput for proposed Approach C is greater than the other approaches. It is also evident that larger is the offloaded data disparity, the larger is the difference between the system's effective throughput for the compared schemes. The reason behind this trend is the optimal core allocation. When the offloaded data is same in characteristic (i.e., amount and complexity is same), the cores allocation for the schemes are same (i.e., equal number of cores for no optimization, Approach A and Approach C) and the difference in the throughput appears only because of the power allocation. However, as the offloaded data disparity increases, the proposed scheme outperforms others. The average increase in the system effective throughput is 19% for the case shown in Fig. 4.

V. Conclusion

In this letter, it has been shown that the transaction time plays an important role in improving the overall resource utilization and the transaction time difference of two users is minimized by optimizing both the transmission powers and computational resources allocation independently. The proposed optimization resulted in increased effective throughput of the system. As a future direction to this work, the joint problem can also be investigated while considering correlation both communication and computation resources. The approach can be extended to multiple users in a NOMA cluster. A data aware NOMA clustering scheme can be used where the users are paired considering both the power disparity as well as their data offloading requirements, which can contribute further towards improvement of spectral efficiency and system's effective throughput.

References

[1] S. Qureshi, S. A. Hassan, and D. N. K. Jayakody, "Divide-and-allocate: An uplink successive bandwidth division NOMA system," *Transactions on Emerging Telecommunications Technologies*, vol. 29, no. 1, p. e3216, 2018.

[2] Y. Wang, M. Sheng, X. Wang, L. Wang, and J. Li, "Mobile-edge computing: Partial computation offloading using dynamic voltage scaling," *IEEE Transactions on Communications*, vol. 64, no. 10, pp. 4268–4282, 2016.

[3] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "Mobile edge computing—a key technology towards 5g," *ETSI white paper*, vol. 11, no. 11, pp. 1–16, 2015.

[4] Z. Ding, D. W. K. Ng, R. Schober, and H. V. Poor, "Delay minimization for NOMA-MEC offloading," *IEEE Signal Processing Letters*, vol. 25, no. 12, pp. 1875–1879, 2018.

[5] A. Kiani and N. Ansari, "Edge computing aware noma for 5g networks," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1299–1306, 2018.

[6] Z. Yang, J. Hou, and M. Shikh-Bahaei, "Energy efficient resource allocation for mobile-edge computation networks with noma," in *2018 IEEE Globecom Workshops (GC Wkshps)*, pp. 1–7, IEEE, 2018.

[7] Y. Pan, M. Chen, Z. Yang, N. Huang, and M. Shikh-Bahaei, "Energy-efficient noma-based mobile edge computing offloading," *IEEE Communications Letters*, vol. 23, no. 2, pp. 310–313, 2019.

[8] Z. Ding, J. Xu, O. A. Dobre, and V. Poor, "Joint power and time allocation for noma-mec offloading," *IEEE Transactions on Vehicular Technology*, 2019.

[9] Z. Yang, J. Hou, and M. Shikh-Bahaei, "Resource allocation in full-duplex mobile-edge computing systems with noma and energy harvesting," *arXiv preprint arXiv:1807.11846*, 2018.

[10] M. S. Ali, H. Tabassum, and E. Hossain, "Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (noma) systems," *IEEE access*, vol. 4, pp. 6325–6343, 2016.

[11] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[12] J. Cui, Y. Liu, Z. Ding, P. Fan, and A. Nallanathan, "Qoe-based resource allocation for multi-cell noma networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 9, pp. 6160–6176, 2018.

[13] J. Cui, Y. Liu, Z. Ding, P. Fan, and A. Nallanathan, "Optimal user scheduling and power allocation for millimeter wave noma systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 3, pp. 1502–1517, 2017.