Cross-Talk: D2D Potentiality Based Resource Borrowing Schema for Ultra-Low Latency Transmission in Cellular Network

Guolin Sun, Timothy Dingana, Sebakara Samuel Rene Adolphe and Gordon Owusu Boateng
School of Computer Science and Engineering, University of Electronic Science and Technology of China
Chengdu-China
[e-mail: guolin.sun@uestc.edu.cn]
*Corresponding author: Guolin Sun

Received July 31, 2018; revised September 28, 2018; revised November 7, 2018; accepted November 16, 2018; published May 31, 2019

Abstract

Resource sharing is one of the main goals achieved by network virtualization technology to enhance network resource utilization and enable resource customization. Though resource sharing can improve network efficiency by accommodating various users in a network, limited infrastructure capacity is still a challenge to ultra-low latency service operators. In this paper, we propose an inter-slice resource borrowing schema based on the device-to-device (D2D) potentiality especially for ultra-low latency transmission in cellular networks. An extended and modified Kuhn-Munkres bipartite matching algorithm is developed to optimally achieve inter-slice resource borrowing. Simulation results show that, proper D2D user matching can be achieved, satisfying ultra-low latency (ULL) users' quality of service (QoS) requirements and resource utilization in various scenarios.

Keywords: device-to-device communication, resource cross-borrowing, bipartite matching, ultra-low latency

http://doi.org/10.3837/tiis.2019.05.002 **ISSN : 1976-7277**

1. Introduction

Achieving ultra-low latency (ULL) automation to enhance the sensory and processing capabilities of human beings has been regarded as one of the ultimate goals of the tactile internet. The ULL services in 5G could embrace all upcoming applications, such as unmanned or remote control, augmented reality, intelligent transportation systems, smart grid and the internet of things (IoTs). Therefore, how to improve the user experience of ULL users is an enormous challenge when network congestion occurs.

Resource virtualization techniques and dynamic air-interface slicing are proposed to guarantee various latency levels in order to provide differentiated services for slices [1]. These approaches are optimized for virtual radio resource allocation to enhance the security of mobile social networks in the air interface and are perfectly applicable in device-to-device (D2D) communications for mobile cellular networks due to the limited bandwidth resource [2]. The motivation of this paper is that, we propose a joint buffer and bandwidth management scheme to further reduce queuing delays for ULL flows in a fixed purpose IoT-based ULL service scenario, such as smart homing, smart factory and smart stadium. In [2], a trusted group in social networks was considered for resource sharing. However, the exchange of heterogeneous resources were not considered. In our scheme in this paper, ULL slice which is delay-sensitive temporarily borrows bandwidth resource from best-effort (BE) slice and lends buffer resource to BE slice, which is not sensitive to delay. In this paper, ULL is a representative type of delay-sensitive flows, and BE is a representative type of delay-elastic flows. The aim of D2D potentiality based bipartite matching is to meet the quality of service (QoS) requirements of delay sensitive traffics, improve their throughput through resource cross-borrowing and also to manage network resource efficiently. Despite the enormous advantages of D2D communications, secure data sharing has always been a concern for mobie operators. Recently, some works have focused on the secure data transfer between mobiles in D2D mode. The author in [3] proposed a secure data sharing protocol which merges the advantages of public key cryptography and symmetric encryption, to achieve data security in D2D communication. The term resource in this paper is referred to as the effective bandwidth and buffer needed by ULL users and BE users respectively. Since ULL users are sensitive to delay, enough bandwidth is always needed for data transmission. BE users, who are not sensitive to delay may possess bandwidth resource in abundance which would be needed by the ULL users for their transmission. In this scenario, the ULL slice decides to trade its buffer with bandwidth from the BE slice. How the delay-sensitive slice borrows resources from delay-elastic traffic slice and vice versa is an issue. The resource virtualization and the extended and modified Kuhn-Munkres(KM) based bi-partite matching algorithm are proposed to achieve the optimal node pairing results for inter-slice resource cross-borrowing.

The bandwidth and buffer resources are virtualized and partitioned into two slices: the ULL slice and BE slice. Based on the Kuhn-Munkres bi-partite matching algorithm proposed in [4], users from the ULL slice who are in need of extra bandwidth for data transmission request to match with users in the BE slice who need extra buffer for data storage provided they can communicate in D2D mode. A D2D communication can be defined as the direct data exchange between any two mobile users via a D2D link. The criteria for D2D communication is the satisfaction of signal-to-interference-plus-noise-ratio(SINR) constraint, the maximum potentiality of a ULL user to match with a BE user and minimum matching cost between the said users. The users of BE slice borrow buffer resources from the ULL slice in exchange, temporarily storing users' data. The main contributions of this paper are listed as follows:

- We develop an effective inter-slice resource cross-borrowing scheme between delay-sensitive flows and delay-elastic flows to satisfy the users' latency requirement when the network is congested by exploring their D2D potentialities.
- We modify and customize the KM algorithm to match ULL and BE users capable of resource cross-borrowing in D2D mode efficiently to reduce matching cost and satisfy the QoS requirements of users.

The rest of paper is organized as follows; In Section 2, we review previous works. A system model is presented in Section 3. The formulated non-linear integer programming problem for resource allocation and the proposed modified KM algorithm are described in the Section 4. Comprehensive performance evaluations are discussed in Section 5. Finally, we conclude this paper in Section 6.

2. Related Work

There have been a lot of researches on guaranteeing ULL transmission with various techniques presented. Recently, the notion of 5G-enabled tactile internet is emerging, which is envisioned to enable the delivery of real-time control and physical haptic experiences in perceived real-time [5][6][7]. A. Aijaz proposed a novel radio resource slicing framework, called Hap-SliceR, which aims to achieve ULL in haptic communications using Q-learning technique in [5]. In [6], the authors proposed an adaptive multiplexer, known as Admux, for tactile internet which integrates visual, auditory and haptic requirements in a statistically optimal manner. M. O. Ernst *et al.* also considered the reduced transmission time interval (TTI) for ULL transmissions in [7]. Specifically, the tactile internet requires a round-trip latency of 1ms. From the perspective of the physical layer, each packet must not exceed a duration of 33μs in order to enable a one-way physical layer transmission of 100μs. The authors in [8] provided an initial analysis on ULL random radio access problem for remote control. By reducing the TTI from 1ms to 100μs without retransmission on the air interface, it realized a latency of 1ms. In [9], the authors proposed an idea of trading a little bandwidth for ULL data transmission in the cloud data center. They concluded that by sacrificing a small amount of bandwidth, average and tail latencies in the data centers could be reduced. The proponents of a KM algorithm-based quality of experience (QoE) aware resource allocation for mixed traffics in heterogeneous networks in [10] ensured maximum resource utilization but failed to consider bandwidth and buffer resource exchange. The authors in [11] proposed an optimized resource allocation method for intra-cluster D2D users based on the KM algorithm. However clusters were created and resources populated for all of the users without exploiting joint bandwidth and tradeoff. Virtual resource allocation was formulated as a joint user association and resource allocation model, which is a convex optimization problem and solved with the alternating direction method of multipliers(ADMM) in [12]. Unfortunately, due to the handover delay and signaling burden introduced, the tight combination with association cannot achieve ultra-low latency. To the best of our knowledge, there is still no research work considering resource allocation for ULL flows by exchanging bandwidth resource with buffer resource based on D2D communication, especially in the scenario of heavy load network traffic.

3. System Model

The system model consists of five components in the network namely; user equipment (UEs), slices which provide services to UEs, software defined networking (SDN) controller, base station (BS) and resource, as illustrated in **Fig. 1**. The wireless network is partitioned into slices (in our case, ULL slice and BE slice) in the form of services and managed by mobile virtual network operators (MVNOs). After network slicing, a slice that has been created is expected to be admitted into the network, using a technique known as slice admission control, where it would be assigned to a MVNO. The SDN controller classifies all received requests from the slices according to their service requirements and assigns a slice to the BS based on the aggregated QoS demand of the slice. Next is UE admission control. At this step, UEs are admitted into the network based on their type of service and their QoS demand. For instance, a UE of on-demand video streaming service is expected to belong to the ULL slice rather than the BE slice. UE admission control and resource allocation for throughput maximization are done by the controller whiles the UEs request resources from the slice they belong to. The BS performs signaling control, incorporating proper D2D bipartite matching as well as resource borrowing schema. Resources are classified into bandwidth resource and buffer resource. We assume that, initial resource allocation to both slices is based on a service-level agreement (SLA) between the MVNOs and the infrastructure provider (InP) that owns the resource.

The BS broadcasts its beacon to all UEs at periodic time intervals. Upon reception of a periodic transmit power command in an uplink scheduling grant, the UE adjusts its transmitting energy per resource element (EPRE) accordingly and the BS obtains the channel gain (h_m) of UEs in the coverage area. After obtaining information about the allocated bandwidth for the sub-channel, the BS calculates its achievable data rate. Through the BS, ULL users and BE users can learn from each other to know the number of UEs in different slices around them as well as the SINR of the potential D2D pairs. At this stage, these information (the number of ULL users and the amount of resources needed for the ULL user to satisfy its QoS demand, D2D information among ULL and BE users) is reported to the BS, which runs the resource borrowing module to decide on which UEs to match for cross-borrowing. The resource borrowing module schedules UEs from a ULL slice taking into consideration their required QoS, especially for the users' request with short remaining life time and the number of potential BE users around them. Finally, the resource borrowing module decides on the optimal matching of ULL users to BE users, which is capable of minimizing the matching cost and then relays the decision to the BS. Upon receiving the result of the D2D matching from the resource borrowing module, the BS also assigns the borrowed bandwidth from BE users to their matched ULL users and causes all matched BE users to switch to D2D mode and forwards their packets into ULL users' buffer.

The system model is categorized into business model, network model, virtualization model and utility model. In the business model, we assume the InP owns and leases physical infrastructure and virtualized resources, e.g. bandwidth resource and buffer resource, to the MVNOs at a cost. MVNOs provide services to UEs with the virtualized resources they acquired from the InP. The UEs can subscribe to these services at a fixed cost per month. Resources (bandwidth and buffer) allocated to a MVNO by the InP are shared amongst its UEs. In case the portion of bandwidth resource allocated to ULL slice cannot satisfy the QoS requirements of its UEs, the ULL slice can borrow bandwidth resource from the BE slice on the condition that the ULL slice exchanges with an equal buffer size. In the case of a heavy-load network, the portion of bandwidth allocated to each UE in the ULL slice may not be enough to meet its transmission requirements which may cause delay. On the other hand, in the case of a heavy or light load network, the portion of bandwidth allocated to each UE in the BE slice may be more than enough to satisfy its transmission requirements for the reason that their packets can afford a moderated delay in their buffers. Although, ULL flows require real-time transmission, they are made up of small packets i.e. their buffers will stay almost empty. BE flows can afford delay; however, packet generation will take place and may exceed their original allocated buffer. With these assumptions, we propose a resource borrowing scheme where ULL UEs borrow bandwidth from BE UEs, and in turn provide to them buffer to avoid packet drops that may occur in case of prolonged delay.

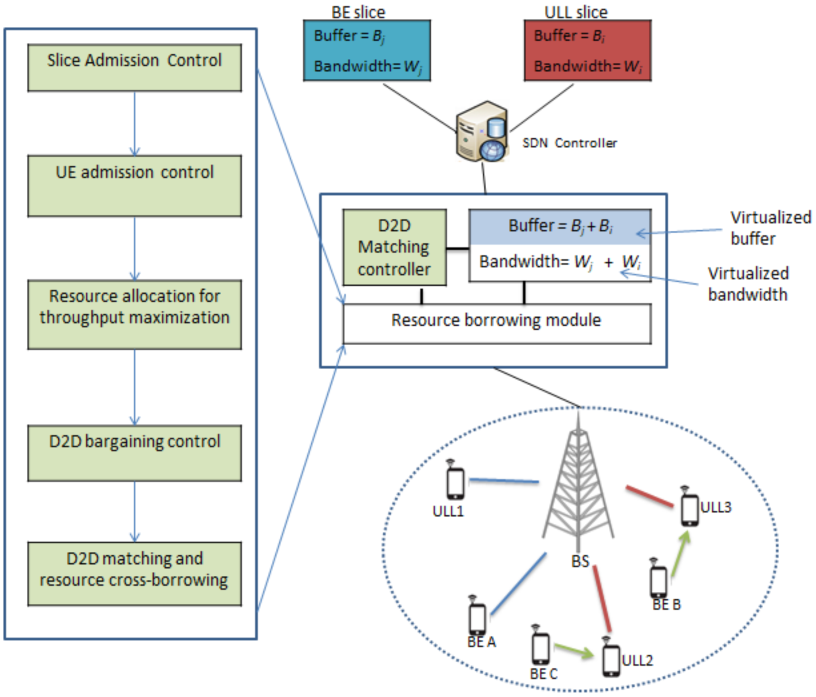


Fig. 1. System model

Let k ∈ Ꝁ $= \{1, 2 ... K\}$ denote a set of UEs in the network and $m \in \mathcal{M} = \{1, 2 ... M\}$ the BS which is linked to a SDN controller. We classify UEs into two slices, preferably BE slice and ULL slice. We indicate i ∈ Ɨ $= \{1, 2 ... I\}$ as a set of UEs that belong to the ULL slice, with j ∈ Ɉ $= \{1, 2 ... J\}$ as the set of UEs that belong to the BE slice, i.e. Ɨ ∪ Ɉ = Ꝁ. We assume that, each UE has a dual transmission mode i.e. both cellular mode and D2D mode. The selection of the transmission mode depends on the application scenario. For resource borrowing between a ULL user and a BE user, eligibility of D2D resource reuse is considered. Assmuing that C orthogonal resource blocks are available in a cell, the BS allocates resources to C cellular users. Moreover, only one or zero D2D user is allowed to share the same resource with the cellular user. Let $d \in \mathcal{D} = \{1,2, ..., D\}$ represent a set of D2D users and $c \in \mathcal{C} = \{1,2, ..., C\}$ be a set of cellular users in the network. Each slice $s \in \hat{S} = \{1, 2 ... S\}$ has a priority Θ_s based on the type of service it provides and is initially allocated resources based on the SLA between the operator and the InP. A slice is allocated resource in the form of bandwidth and buffer. Each slice's demand is determined by the constraint $Rs\ (W_s, Td_s)$, where W_s denotes the bandwidth requirement of the slice and Td_s denotes its affordable time delay. Each UE k is characterized by the throughput demand T_k, which enables the slice to allocate to it a portion of its bandwidth. The amount of bandwidth initially allocated to UE k from slice s is expressed as;

$$\mathrm{w}_k = \frac{W_s}{n_s}, \tag{1}$$

where n_s is the total number of UEs in the slices s to which the UE k belongs and W_s is the bandwidth allocated to slice s.

3.1 Bandwidth-based virtualization model

We compute the initial achievable data rate of UE i in the ULL slice on BS m as

$$\acute{\Gamma}_{im} = w_{im}.\, log_2(1 + \bar{\text{ṛ}}_{im}), \tag{2}$$

where w_{im} denotes the amount of bandwidth allocated to UE i from BS m. $\bar{\text{ṛ}}_{im}$ denotes the SINR of UE i with BS m which is defined as;

$$\bar{\text{ṛ}}_{im} = \frac{h_{im}.P_m}{N_o + h_{in,n\neq m}.P_n}, \tag{3}$$

where P_m and h_{im} denote the transmit power of BS m and the channel gain between UE i and BS m respectively. P_n is the received interference power at the other BSs except the BS m, h_{in} is the channel gain between the BS n and UE i. N_o is noise spectral density. Based on equation (2), we can calculate the extra bandwidth needed by UE i for data transmission given its bandwidth demand as follows;

$$\Delta \mathrm{w}_i = w_{di} - w_{im}, \tag{4}$$

where w_{di} is the desired bandwidth or rate and w_{im} is the amount of bandwidth allocated to UE $i$from its serving BS m before resource borrowing. The extra bandwidth ($\Delta \mathrm{w}_i$) can be considered as the additional bandwidth needed by the ULL user i to achieve its desired throughput. To attain this throughput, an additional bandwidth ($\Delta \mathrm{w}_i$) has to be borrowed from the BE slice. The data rate of UE i after borrowing extra bandwidth from a BE user can be expressed as;

$$\Gamma'_{im} = (\mathrm{w}_{im} + \Delta \mathrm{w}_i).\, log_2(1 + \bar{\text{ṛ}}_{im}). \tag{5}$$

3.2 Buffer-based virtualization model

Let the total buffer size allocated to UE k be B_k, the occupied buffer size being b_k and Δb_k denoting the remaining buffer size. The remaining buffer size of UE k can be expressed as;

$$\Delta b_k = B_k - b_k, \tag{6}$$

The total virtualized buffer of the network is made up of the ULL slice's buffer B_i and the BE slice's buffer B_j. The total virtualized buffer on the BS m, B_m is computed as;

$$B_m = B_i + B_j. \tag{7}$$

When two UEs from different slices intend to cross-borrow resources from each other, a D2D link is established between them. After the resource cross-borrowing, the BE user obtains additional buffer from the ULL user and the total BE buffer is expressed as;

$$B_{jm} = \propto B_i + B_j, \tag{8}$$

where $\propto B_i$ is the fraction of the ULL slice's buffer that is obtained as a result of the exchange.

3.3 D2D Potentiality and matching cost model

We define D2D potentiality as the probability or magnitude of chance of two UEs of different slices (ULL slice and BE slice) to engage in resource cross-borrowing. Each UE i, has a potentiality Q_{ij} as the probability of matching with an intended UE j. Before cross-borrowing, there is bargaining amongst a number of ULL users who wish to trade buffer with bandwidth and BE users who are ready to do otherwise. One ULL user has a set of BE users known as BE bargainers who compete with themselves to match with the ULL user. We represent the set of BE bargainers as Ƶ_i and the number of bargainers in the set is denoted as $|\text{Ƶ}_i|$. Bargaining is done to assess whether the satisfaction-based and matching-based D2D constraints are met for the intended D2D pair. We define Q_{ij} as the potentiality of one ULL user i to match with a UE j in the set of BE bargainers Ƶ_i as;

$$\mathrm{Q}_{ij} = \frac{1}{|\text{Ƶ}_i|} \tag{9}$$

A user j will be selected as the winner bargainer by user i, and a proportion of bandwidth of user j is allocated to user i. A D2D link will be made up of two UEs who must be in satisfaction relationship i.e. the aggregated data to be transmitted or stored by a UE should be either satisfied by the aggregated throughput Γ'_{im} in the case of UE i or the aggregated buffer B_{jm} in the case of UE j respectively. From data rate demand, we define the amount of data needed to be transmitted by each UE in a given time period tp as $d_i = T_i.tp$, where d_i denotes the data packets in bits generated in a tp time period. For satisfaction, l_{ij} = 1 if the aggregated buffer B_{jm} can store any remaining data that was not able to be transmitted at a given time period tp, otherwise l_{ij} = 0.

$$l_{ij} = \begin{cases} 1, & B_{jm} \geq \left(\Gamma'_{im} - \left(T_i + T_j\right)\right).tp \\ 0, & B_{jm} < \left(\Gamma'_{im} - \left(T_i + T_j\right)\right).tp \end{cases} \tag{10}$$

The potentiality of a D2D link can also be computed from the potentialities Q_{ij} of UE i to UE j and Q_{ji} of UE j to UE i, who intend to exchange resource with each other as follows;

$$\prod_{ij} = \frac{\mathrm{Q}_{ij}+Q_{ji}}{2} \times l_{ij} \tag{11}$$

In order for two UEs to be in D2D mode, the SINR of the potential D2D link ṝ_{ij} must exceed the minimum SINR threshold ṝ_{ij}^{min} as;

$$\text{ṝ}_{ij} = \frac{h_{ij}.P_{ij}}{N_o + h_{im}.P_m} \geq \text{ṝ}_{ij}^{min} \tag{12}$$

where P_{ij} denotes the transmit power of the D2D transmitter and h_{ij} is the D2D channel gain. $h_{im}.P_m$ is the interference power from BS m. From equation (11) and equation (12), we formulate our matching cost as:

$$\phi_{ij} = \text{ṝ}_{ij} \times \prod_{ij} \tag{13}$$

where ϕ_{ij} denotes the matching cost. Finally, we define the utility models in the system model for performance metrics to evaluate the proposed schema.

3.4 ULL QoS on throughput

We first compute the aggregate throughput of the ULL slice before and after resource cross-borrowing in equation (14) and (15) respectively as;

$$T\_ull_{agg1} = \sum_{i=1}^{I} \Gamma_{im}\,,\ i \in \text{Ɨ} \tag{14}$$

$$T\_ull_{agg2} = \sum_{i=1}^{I} \Gamma'_{im}\,, i \in \text{Ɨ} \tag{15}$$

We define the ULL QoS in terms of satisfaction T_ull_{sat}, as a ratio of the aggregate throughput after resource cross-borrowing to the aggregate throughput before resource cross-borrowing as;

$$T\_ull_{sat} = \frac{T\_ull_{agg2}}{T\_ull_{agg1}} \tag{16}$$

3.5 Latency Satisfaction

For slice latency satisfaction, we define a UE's latency as the time used to transmit its data demand using the achievable data rate.We calculate the latency after cross-borrowing for ULL user L_ull_{ij} and BE user L_be_{ij} in equations (17) and (18) respectively as;

$$L\_ull_{ij} = \frac{T_i}{\Gamma'_{im}} \tag{17}$$

$$L\_be_{ij} = \frac{T_j}{\Gamma'_{jm}} + L\_ull_{ij} \tag{18}$$

where L_ull_{ij} denotes the latency of ULL user i matched with BE user j, and L_be_{ij} denotes the latency of BE user j matched with ULL user i.

After obtaining the user latency, we can easily define the slice's latency as the aggregate of its users' latencies as follows;

$$L_{\_ull} = \sum_{i=1}^{I} L\_ull_{ij} \tag{19}$$

$$L_{\_be} = \sum_{j=1}^{J} L\_be_{ij} \tag{20}$$

where $L_{\_ull}$ and $L_{\_be}$ denote ULL slice's latency and BE slice's latency respectively.

3.6 Algorithm fairness

We define the algorithm's matching fairness as the ratio of matched ULL users to the total number of ULL users within the system. We define M_{ij} as the number of matched ULL users, and we compute the matching fairness as:

$$F_y = \frac{M_{ij}}{\mathrm{I}}, i \in\ \text{Ɨ} \tag{21}$$

4. Problem Formulation

In this section, we aim to maximize the throughput gain of UEs first and then propose a solution to find a good match, satisfying the QoS requirements of the UEs of ULL slice and BE slice and minimizing their matching cost. Firstly, we formulate the physical resource allocation problem for cellular users and D2D users based on throughput maximization. On the other hand, we formulate the inter-slice resource cross-borrowing problem as a non-linear integer programming problem to minimize matching cost. The detailed description of solving the two optimization problems are presented as follows.

4.1 Resource allocation for throughput maximization

In order to maximize the overall network throughput, we first determine the subset of D2D users that can access the resource of cellular users (D_A) considering slice resource, user QoS and transmit power constraints. It should be noted that, the slice resource constraint is different for different types of slices. By solving the objective function, the optimal power and resource allocation and maximum throughput can be obtained. Considering that cellular and D2D links co-exist in the cell, the physical resource allocation problem can be formulated, also as an extension of the formulation in [13]:

$$max_{x_{(c_s,d_s)},P_{c_s},P_{d_s}} \sum_{s=1}^{S} \sum_{c_s \in C} \sum_{d_s \in D} \left\{ w_{c_s m} \cdot log_2\left(1 + \bar{ṛ}_{c_s m}\right) + x_{(c_s,d_s)} \cdot w_{d_s m} \cdot log_2(1 + \bar{ṛ}_{d_s m}) \right\} \quad (22)$$

such than;

$$\bar{ṛ}_{c_s m} = \frac{P_{c_s} \cdot h_{c_s,m}}{N_o + P_{d_s} \cdot h_{d_s,m}} \geq \bar{ṛ}_{c_s m}^{min} \; ; \quad (23)$$

$$\bar{ṛ}_{d_s m} = \frac{P_{d_s} \cdot h_{d_s,m}}{N_o + P_{c_s} \cdot h_{c_s,m}} \geq \bar{ṛ}_{d_s m}^{min} \; ; \quad (24)$$

$$\sum_{d_s} x_{(c_s,d_s)} \leq 1, \quad x_{(c_s,d_s)} \in \{0,1\} \; ; \quad (25)$$

$$\sum_{c_s} x_{(c_s,d_s)} \leq 1, \quad x_{(c_s,d_s)} \in \{0,1\} \; ; \quad \forall c_s \in C; \; \forall d_s \in D_A \quad (26)$$

$$0 \leq P_{c_s} \leq P_{c_s}^{max} \; ; \quad (27)$$

$$0 \leq P_{d_s} \leq P_{d_s}^{max} \; ; \quad \forall c_s \in C; \; \forall d_s \in D_A \quad (28)$$

$$\sum_{s=1}^{S} \sum_{c_s \in C} w_{c_s m} \leq w_{sm}^{max} \quad \forall s \in S; \quad (29)$$

where P_c and P_d indicate the transmission power of cellular user and D2D user respectively, $D_A (D_A \in D)$ represents the subset of D2D users that can access the cellular network, $\bar{ṛ}_{c_s m}^{min}$ and $\bar{ṛ}_{d_s m}^{min}$ are the minimum SINR threshold for cellular users and D2D users respectively. $x_{(c_s,d_s)}$ is the resource reuse identifier where $x_{(c_s,d_s)} = 1$ means the D2D users d reuses the resource of cellular user c and $x_{(c_s,d_s)} = 0$, otherwise. Constraint (29) ensures that, the sum of the bandwidth resource of the cellular users in a slice should not exceed the maximum bandwidth resource allocated to the slice. Then, we propose the resource cross-borrowing scheme for the ULL slice and the BE slice with the constraint of slice resource. For simplicity, we consider two particular slices, preferably BE slice and ULL slice. In the next section, we denote $i \in Ɨ = \{1, 2 ... I\}$ as a set of UEs that belong to the ULL slice, and $j \in Ɉ = \{1, 2 ... J\}$ as the set of UEs that belong to the BE slice.

4.2 Inter-slice resource cross-borrowing

In this section, we formulate the inter-slice resource cross-borrowing problem as a non-linear integer programming problem and propose a solution to find a good match, satisfying the QoS requirements of UEs of both slices, and minimizing their matching cost. The resource cross-borrowing problem can be formulated as;

$$C = min \sum_{i=1}^{I} \sum_{j=1}^{J} -(\phi_{ij} \, . \, \mathrm{a}_{ij}) \quad (30)$$

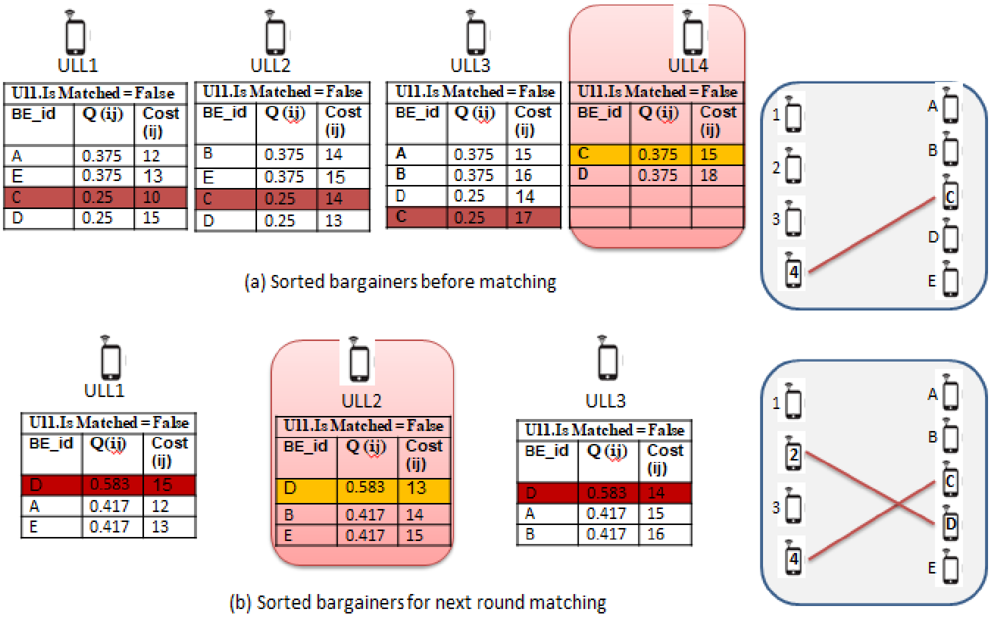
such that;

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \mathrm{a}_{ij} = 1 \quad (31)$$

$$\sum_{i=1}^{I} \sum_{j=1}^{J} l_{ij} \; . \; \mathrm{a}_{ij} = 1 \quad (32)$$

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\bar{\mathrm{r}}_{ij}\,.\,\mathrm{a}_{ij}\ \leq \bar{\mathrm{r}}_{ij}^{min} \tag{33}$$

where a_{ij} is the matching indicator with binary value 1 or 0. Constraint (31) states that, a UE of the ULL slice can match with only one UE of the BE slice and vice versa. Constraint (32) ensures that both UEs who intend to cross-borrow resources in D2D mode must be in a satisfaction relationship. Constraint (33) shows that the SINR of the D2D link between the intended UEs must exceed the minimum SINR threshold. However, the formulated problem is NP-hard by nature. With such a problem, the only solution could be an approximation to the optimal solution. Therefore, we propose a low-complexity heuristic algorithm to solve the matching problem accurately and efficiently. Based on the KM algorithm, we propose an extended and modified KM bipartite matching algorithm to solve the above formulated problem. With the KM algorithm, there is equal number of UEs on both sides for matching. Unlike the KM algorithm, the extended and modified KM bipartite matching algorithm has unequal number of ULL users and BE users.



ULL1

Ull.Is Matched = False		
BE_id	Q (ij)	Cost (ij)
A	0.375	12
E	0.375	13
C	0.25	10
D	0.25	15

ULL2

Ull.Is Matched = False		
BE_id	Q (ij)	Cost (ij)
B	0.375	14
E	0.375	15
C	0.25	14
D	0.25	13

ULL3

Ull.Is Matched = False		
BE_id	Q (ij)	Cost (ij)
A	0.375	15
B	0.375	16
D	0.25	14
C	0.25	17

ULL4

Ull.Is Matched = False		
BE_id	Q (ij)	Cost (ij)
C	0.375	15
D	0.375	18

(a) Sorted bargainers before matching

ULL1

Ull.Is Matched = False		
BE_id	Q(ij)	Cost (ij)
D	0.583	15
A	0.417	12
E	0.417	13

ULL2

Ull.Is Matched = False		
BE_id	Q (ij)	Cost (ij)
D	0.583	13
B	0.417	14
E	0.417	15

ULL3

Ull.Is Matched = False		
BE_id	Q (ij)	Cost (ij)
D	0.583	14
A	0.417	15
B	0.417	16

(b) Sorted bargainers for next round matching

Fig. 2. The extended and modified KM sketch

As an example, as shown in **Fig. 2**, there are four UEs of the ULL slice (ULL1, ULL2, ULL3 and ULL4) and five UEs of the BE slice (*A*, *B*, *C*, *D* and *E*). For each of the ULL users, there are BE users who have the potentiality to match and cross-borrow resources with them. From **Fig. 2**, each possible match has a potentiality (Q_{ij}) and a matching cost ($Cost_{ij}$). The values of the potentiality and matching cost were generated using MATLAB software. We first calculate the extra amount of bandwidth needed to be borrowed by the ULL user from a BE user to satisfy their QoS requirements. The resource borrowing module tends to find the BE users who satisfy the conditions for resource borrowing for each ULL user. These BE users are referred to as the set of bargainers. The set of bargainers for ULL1 are *A*, *E*, *C* and *D* and that of ULL4 are *C* and *D*. Next, all the ULL users are sorted based on the size of their sets of bargainers. The BE user that is common to all ULL users is considered first for matching. The ULL user with the least number of bargainers is then selected for matching with the bargainer with the highest potentiality and minimum matching cost. In a case where the highest potentiality of two or more matches is equal, the best match is selected based on the minimum matching cost and vice versa. From **Fig. 2(a)**, ULL4 has the least number of bargainers. The selected ULL user is matched with the bargainer who satisfies the D2D satisfaction relationship and minimum matching cost amongst other bargainers in that set. Here, ULL4 is matched with BE C since ULL4-BE C matching has a potentiality of 0.375 and a matching cost of 15. Finally, the table of possible matches is updated by removing the matched ULL and BE users from the bargainers' set. From **Fig. 2(b)**, it can be observed that ULL4 and BE C have been removed leaving ULL1, ULL2 and ULL3 to match with the rest of the BE users. The potentialities and matching costs of the possible matches are again generated. BE D is selected for matching because it is the UE that is common to all the ULL users. Here, ULL2 is matched with BE D since ULL2-BE D matching has a potentiality of 0.583 and a minimum matching cost of 13 among the other possible matches. The process is repeated until all ULL users are matched with possible BE users.

Algorithm: The extended and modified Kuhn-Munkres algorithm for resources cross-sharing

```
Step1: Initialization [M ($P_m$, $W_m$), K ($P_k$, $B_k$, $w_k$, $T_k$, $Td_k$)];
define I, J = [];
For each time window
  Syst.Queue.add (K);
  classify K into slices;
  I = I. append ($K_i$) ∀ users of ULL slices;
  J = J. append ($K_j$) ∀ users of BE slices;
End
step2: generate BE bargainers' sets for all users
  Z = []
For each i in I
    calculate $\Delta w_i$ using equation (4);
    For each j in J
        calculate $\bar{\underset{.}{r}}_{ij}$ using equation (12);
        calculate $l_{ij}$ using equation (10);
       If $\bar{\underset{.}{r}}_{ij}$ >= $\bar{\underset{.}{r}}_{ij}^{min}$ AND $l_{ij}$ == 1
         calculate $\phi_{ij}$using equation (13);
         Z.append (i, j, $\bar{\underset{.}{r}}_{ij,}\phi_{ij}$);
       End
   End
End
step3: find proper match
Temp (i) = [];
For each i ∈ I
   generate table $Z_{ij}$ from Z;
     /* sub set of all possible BE bargainers of i.
  calculate $\prod_{ij}$ using equation (11), ∀ j ∈ $Z_{ij}$;
  $Z_{ij}$ .add ($\prod_{ij}$), ∀ i, j ∈ $Z_{ij}$;
  Temp (i) = $Z_{ij}$.sort ($\prod_{ij,}$, $\phi_{ij}$);
End
Z.sort (i, $\prod_{ij,}$);
While Z. IsEmpty () == false
  If i.ismatched==false
     $a_{ij}$ = match (I, i. $Z_{ij}(1,2)$) ;
     i.ismatched==True;
    $Z_{ij}$ .remove(j) ∀ i ∈ Z ;
    GoTo Step3
  End
End
```

The extended and modified KM bipartite matching algorithm description is summarized below: In Step 1, a set of parameters about the BS and UEs is initialized. These parameters are; the BS's transmit power P_m, the total bandwidth at the BS W_m, UE's transmit power P_k, the total bandwidth allocated to UEs B_k, the bandwidth allocated to a UE w_k, throughput demand of the a UE T_k and its affordable time delay Td_k. We assume that, the system receives all UEs in a predefined time window. From lines 3~7, the classifier collects all the UEs and classifies them into service-based slices in the form of ULL and BE. UEs are distributed in the slices' queues as classified before and each slice's waiting queue is updated. In Step 2, we generate sets of BE bargainers for all ULL users by calculating the number of UEs from each slice willing to cross-borrow resources. From lines 11~16, we calculate the extra amount of bandwidth needed to be borrowed by the ULL user from a BE user to satisfy their QoS requirements using equation (4). We use equations (12) and (10) to check the feasibility of D2D communication and satisfaction relationship respectively. In lines 17~18, we calculate the matching cost using equation (13). The bargainers' set of each D2D link is then updated. In Step 3, we find the proper match for optimal resource cross-borrowing. From lines 24~26, we use equation (11) to calculate the full D2D potentiality and update all sets of bargainers for all ULL users. From line 27~38, we sort bargainers for each set by potentiality and matching cost. We also match ULL users with high D2D potentiality with one of its BE bargainers yielding minimum matching cost. In line 35, we remove the matched users from their slices' queues and repeat step 3 until every ULL user has got its proper match to cross-borrow resource.

5. Results and Analysis

To evaluate the performance of our proposed algorithm, we perform numerical simulations using MATLAB. We describe the simulation parameters used for performance benchmarking in **Table 1**.

Table 1. Simulation Parameters

Parameter and units	**Configuration**
Number of BSs	1
Number of ULL users	10-50 (Random)
Number of BE users	50 (Random)
Number of slices	2
Radius of BS coverage	300 meters
Bandwidth of BS	10 MHz
D2D link's path-loss model	148 + 40 log10 d
Cellular link's Path-Loss model	128.1 + 37.6 log10 d
D2D link's transmit power	15 dBm
Cellular link's transmit power	20 dBm
Noise spectral density	−174 dBm/Hz
ULL slice resource	Fixed Fraction, e.g. 8%
BE slice resource	Fixed Fraction, e.g. 70%

The parameters were chosen based on LTE standards. In a given area, a BS, a SDN controller, a set of ULL users and a set of BE users are randomly scattered in the network with a BS coverage of radius 300m. There are 2 services according to which we have 2 slices namely; ULL slice and BE slice. Results are presented in terms of cost minimization, user QoS satisfaction on throughput and delay, model fairness, algorithm accuracy and resource utilization. For the proposed model fairness, we set the number of ULL user range to 10~30 in 60 users in the system, the remaining being BE users. The delay satisfaction experiment is conducted to show the impact of resource cross-borrowing, with 10 to 50 ULL users and the number of BE users fixed to 50. The bandwidth of the BS is set at 10MHz. The transmit power of a D2D link is set at 15dBm and the transmit power of a cellular link is set at 20dBm with a noise spectral density of -174dBm/Hz. The ULL users have their specific throughput requirements and amount of buffer resource. The BE slice users have their allocated bandwidth resource and throughput requirements as well.

In our experiment, we assume the slice characteristics based on LTE standards as follows; For ULL slice the data rate demand varies between 150Kbps to 200Kbps, with the common maximum time delay of 50ms. For BE slice we assume the data rate demand varies between 250Kbps to 380Kbps with its maximum time delay set at 100ms. The proposed algorithm, the potentiality-based (P-based) matching algorithm is evaluated against distance-based (D-based) matching algorithm [14] and random-based (R-based) matching algorithm [15]. With the D-based matching algorithm, the distance derived from the SINR of the D2D link is the main determinant of the matching. R-based matching algorithm on the other hand, considers the matching of UEs in a random manner, which is based on a random function $R_{f(i,j)} \sim \exp(\frac{1-v^2}{4})$ for matching presented in [16] with v denoting the number of vertexes on either side (in our case v =1). In this method, we apply an incremental algorithm to make a greedy search in which the best mapping of the pattern nodes to the data nodes is calculated with the help of an objective function.

5.1 Cost minimization

We perform a simulation on the matching cost yielded by the three matching algorithms: P-based, D-based and R-based. We run the experiment for 50 rounds of simulation to minimize the randomness in results and 10 repetitions for every simulation to guarantee steady results and accuracy.

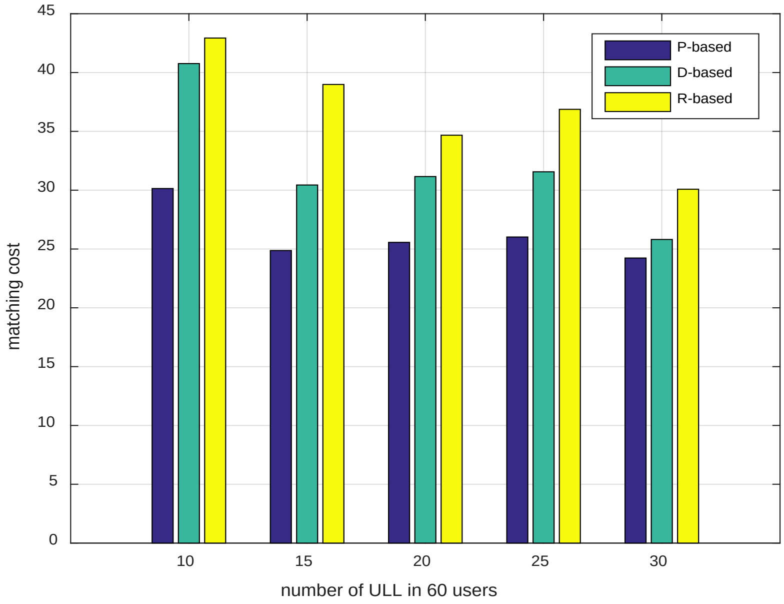


Fig. 3. Cost minimization

As presented in **Fig. 3**, with 10 ULL users, the P-based matching algorithm attains a matching cost of 30 compared with 41 and 43 in the D-based and R-based matching algorithms. As the number of ULL users increase to say 30 ULL users, the matching cost of all three algorithms decrease with the P-based matching algorithm attaining the minimum matching cost of 24 as against 26 and 30 in the other two algorithms. It is clear that, our proposed P-based matching algorithm undoubtedly minimizes the matching cost against D-based matching algorithm and R-based matching algorithm which makes it the best choice for our scenario. It can also be observed that, for all the three algorithms, an increase in the number of ULL users after 10 users decreases the matching cost. Also, different ULL users have different bargaining requirements and potentialities. As the number of ULL users increases, their potentialities and bargaining power change and as a result changing their matching costs.

5.2 QoS Satisfaction

The results shown in **Fig. 4** and **Fig. 5** are the outcome of comparing the three D2D matching algorithms in terms of throughput satisfaction and delay satisfaction respectively. We evaluate the QoS satisfaction in terms of ULL throughput satisfaction ratio as defined in equation (16). Simulation results show that the P-based algorithm outperforms the D-based and the R-based matching algorithms. The satisfaction increases as the number of ULL users increases.

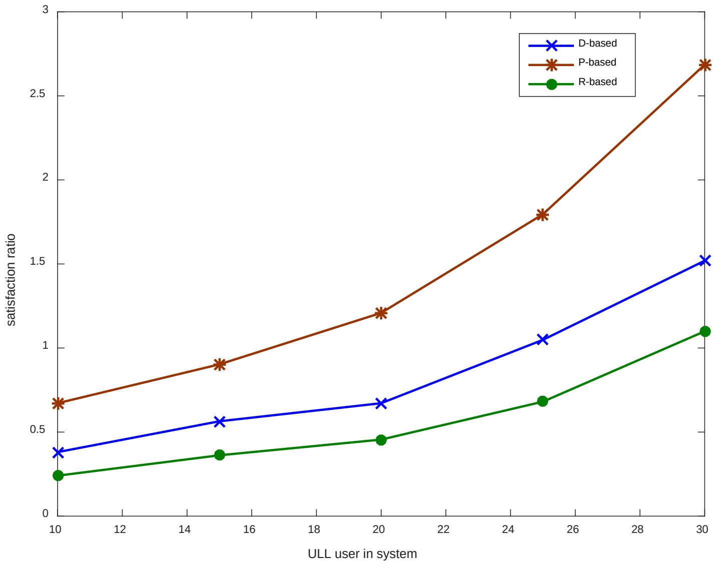


Fig. 4. ULL Throughput satisfaction

In this simulation, we evaluate the QoS satisfaction in terms of latency satisfaction as presented in **Fig. 5**. Here, we consider the effect of resource borrowing by ULL users on the BE users' QoS. With the number of ULL users in the network between 10 and 19, there is no cross-borrowing yet and the latency satisfaction of the UEs is met. As the UEs increase from 20 onwards, the latency of the UEs increase and resource cross-borrowing between ULL users and BE users become necessary. The ULL users exchange bandwidth with buffer with the BE users to satisfy the latency requirements of UEs. As shown in **Fig. 5**, during cross-borrowing the delay of the ULL users is below the ULL maximum delay of 0.05s and that of the BE users is below the BE maximum delay of 0.1s. The experiment results show that the P-based resource borrowing scheme guarantees low latency satisfaction for both ULL and BE users without the distortion of transmission delay requirements of both slices.

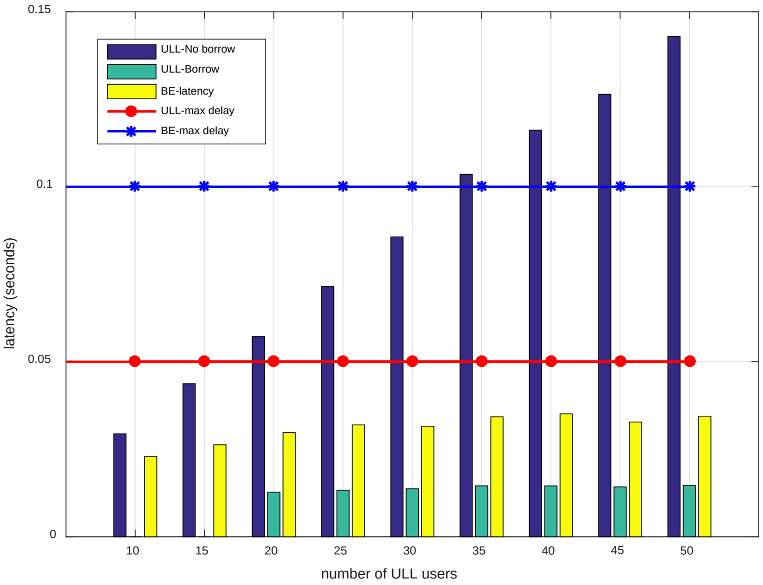


Fig. 5. Delay satisfaction

5.3 Model fairness

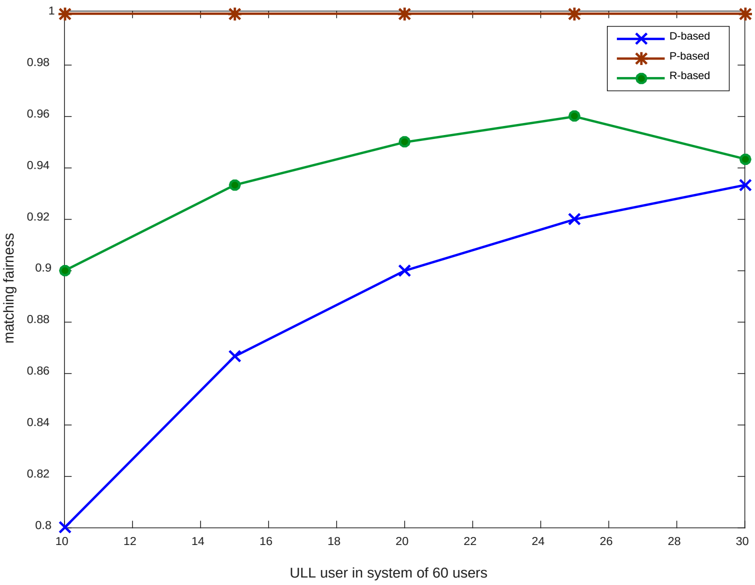


Fig. 6. Model fairness

In this simulation, we compare the model fairness of the three matching algorithms. As shown in **Fig. 6**, our P-based matching algorithm attains a constant fairness of 1.0 as the number of ULL users change which indicates that, it matches all the ULL users to BE users, while the D-based and R-based algorithms attain initial fairness of 0.9 and 0.8 respectively and increase as the number of ULL users increase but are not able to reach a fairness of 1.0, meaning they cannot match all ULL users to BE users. As their fairness coefficient decreases, efficiency also decreases. It can be concluded that, the P-based matching algorithm achieves the best matching fairness which makes it the best matching algorithm for our scenario.

5.4 Algorithm accuracy

In this simulation, the accuracy of our proposed heuristic algorithm in terms of achieving the optimal solution is tested, and the results are presented in **Fig. 7**. As shown in **Fig.7**, the proposed extended and modified KM matching algorithm achieves a lower matching cost compared to the convex-based optimization algorithm with YALIMP solver [17] which is based on the branch and bound algorithm. Moreover, the proposed algorithm is less complex in terms of execution time compared to the convex-based optimization algorithm.

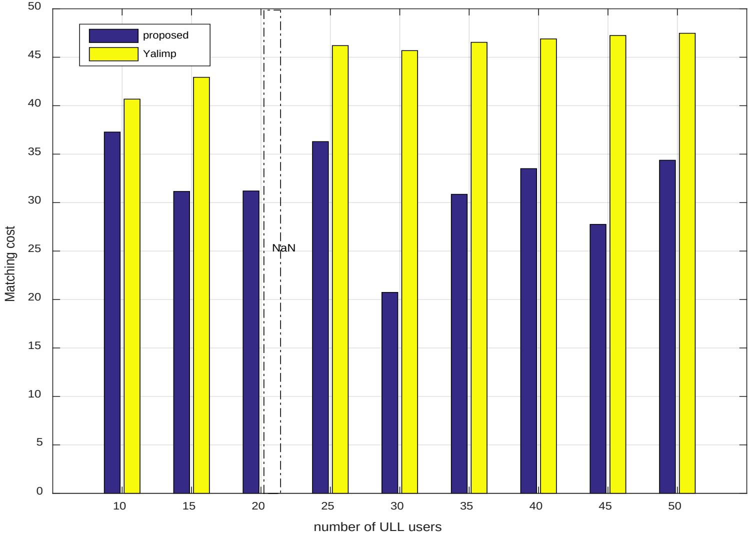


Fig. 7. Algorithm accuracy

5.5 Resource Utilization

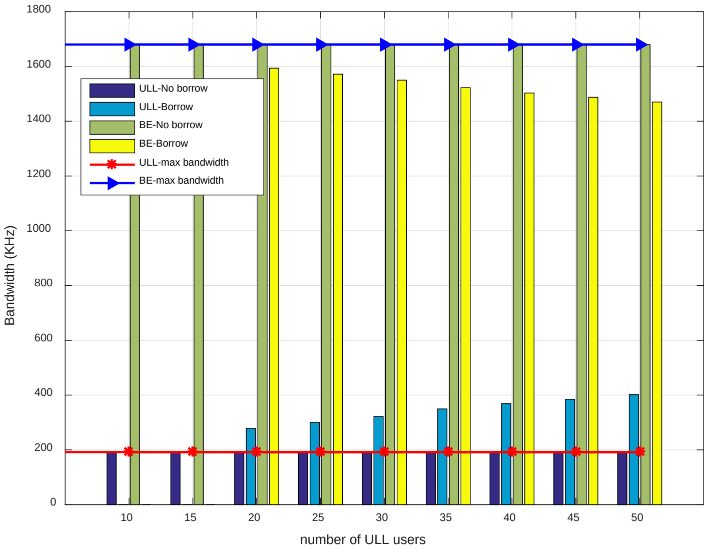


Fig. 8. Bandwidth occupancy

We evaluate the resource utilization in terms of bandwidth and buffer occupancy before and after resource borrowing, for both ULL users and BE users. For bandwidth utilization, we measure how the bandwidth resource is shared and utilized among the two slices' UEs during resource cross-borrowing. As presented in **Fig. 8**, though ULL users borrow almost all of the bandwidth resource of BE users, they occupy only the extra bandwidth needed for their QoS satisfaction which is calculated by equation (5). The remaining bandwidth is used to transmit the BE data buffered in the borrowed ULL buffer, to avoid dropping of packets on the side of BE users and also to avoid the packet disorder that may occur while BE user keeps its cellular transmission during the resource borrowing scenario.

In **Fig. 9**, we use the buffer occupancy to evaluate the impact of resource cross-borrowing on both BE and ULL users' QoS. As shown in the figure above, a small amount of buffer size was used only to keep BE data for a while, but it could not affect the BE's time delay requirement. We observed zero ULL delay as no ULL data buffered, and also that the aggregated buffer could not be exceeded during the cross-borrowing phenomenon.

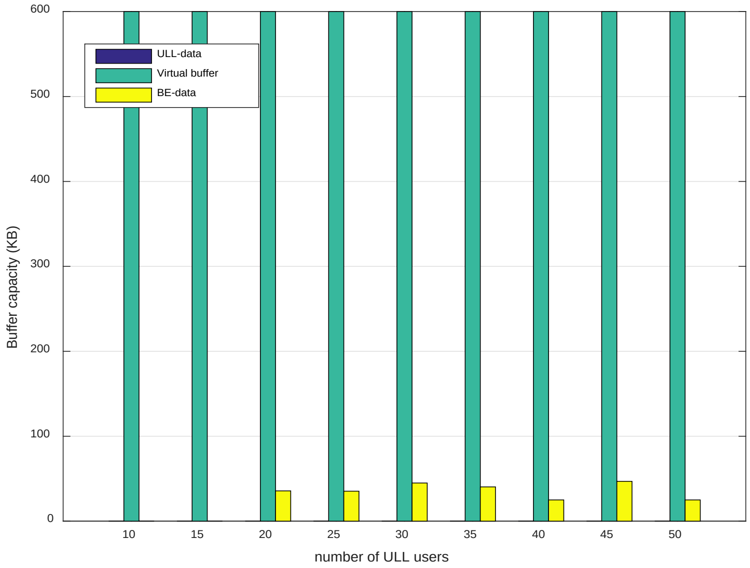


Fig. 9. Buffer Occupany

6. Conclusion

In this paper, we proposed an inter-slice D2D potentiality-based resource cross-borrowing schema for resource allocation in virtualized cellular networks. We formulated the resource allocation problem with a D2D potentiality-based resource borrowing between delay-sensitive flows and delay-elastic flows. The proposed scheme dynamically matches and borrows resources (bandwidth and buffer) among users of different behaviors based on their QoS satisfaction and resource cross-borrowing potentiality in a heavy weighted network. We proposed a modified and extended KM-based heuristic algorithm to perform the D2D matching between slices to ensure QoS satisfaction on heterogeneous characteristics of both slices. The performance evaluation indicated that proper D2D bipartite matching and resource borrowing in various scenarios achieved fairness and inter-slice satisfaction. Moreover, the resource cross-borrowing has a positive impact on 5G where the ULL users will be in need of latency satisfaction in heavily congested network.

Acknowledgment

This work is supported by National Natural Science Research Foundation of China, Grant no. 61771098, by the Fundamental Research Funds for the Central Universities under grant no. ZYGX2018J049, and the ZTE Innovation Research Fund for Universities Program 2016.

References

[1] M. Richart, J Baliosian, J. Serrat and J. L. Gorricho, "Resource slicing in virtual wireless networks: a survey," *IEEE Transactions on Network and Service Management*, vol. 13, no. 3, pp. 462 – 476, September, 2016. Article (CrossRef Link).

[2] Y. He, F. R. Yu, N. Zhao and H. Yin, "Secure social networks in 5G systems with mobile edge computing, caching and device-to-device communications," *IEEE Wireless Communications*, vol. 25, no. 3, pp. 103-109, June 2018. Article (CrossRef Link).

[3] Zhang, Aiqing, et al., "SeDS: Secure data sharing strategy for D2D communication in LTE-Advanced networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 4, pp. 2659-2672, 2016. Article (CrossRef Link).

[4] H. W. Kuhn, "The Hungarian method for the assignment problem, Naval Research Logistic Quarterly," *Naval Research Logistics*, vol. 2, no. 1-2, pp. 83-97, 1955. Article (CrossRef Link).

[5] Adnan Aijaz, "Hap-SliceR: A radio resource slicing framework for 5G networks with haptic communications," *IEEE Systems Journal*, vol. 12, pp. 3, pp. 2285-2296, 2018. Article (CrossRef Link).

[6] M. Eid, J. Cha, and A. El Saddik, "Admux: an adaptive multiplexer for Haptic-audio-visual data communication," *IEEE Trans. on Instrum. and Meas.*, vol. 60, no. 1, pp. 21–31, 2011. Article (CrossRef Link).

[7] M. O. Ernst and M. S. Banks, "Humans integrate visual and haptic information in a statistically optimal fashion," *Nature*, vol. 415, no. 6870, pp. 429–423, January, 2002. Article (CrossRef Link).

[8] Johansson. N. A., Wang Y. P. E., Eriksson E. et al, "Radio access for ultra-reliable and low-latency 5G communications," in *Proc. of IEEE International Conference on Communication Workshop (ICCW)*, pp.1184-1189, 2015. Article (CrossRef Link).

[9] M. Alizadeh, A. Kabbani,T. Edsall, et al. "Less is More : trading a little bandwidth for ultra-low latency in the data center," in *Proc. of NSDI'12 Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*, pp. 253-266, 2012.

[10] N. Wang,Z. Fei,and J. Kuang, "QoE-aware resource allocation for mixed traffics in heterogeneous networks based on Kuhn-Munkres algorithm," in *Proc. of IEEE International Conference on Communication Systems (ICCS)*, pp. 1-6, December 2016. Article (CrossRef Link).

[11] N. Chen, H. Tian and Z. Wang, "Resource allocation for intra-cluster D2D communications based on Kuhn-Munkres algorithm," in *Proc. of IEEE 80th Vehicular Technology Conference (VTC Fall)*, pp. 1-5, September, 2014. Article (CrossRef Link).

[12] C. Liang and F. R. Yu, "Distributed resource allocation in virtualized wireless cellular networks based on ADMM," in *Proc. of Proceedings of IEEE INFOCOM 2015*, pp. 360-365, August, 2015. Article (CrossRef Link).

[13] Y. Wu, X. Liu, X. He, Q. Yu and W. Xu, "Maximizing throughput gain via resource allocation in D2D communications," *EURASIP Journal on Wireless Communications and Networking*, no. 1, pp.1-9 , December 2017. Article (CrossRef Link).

[14] M. C. Lucas-Estan and J. Gozalvez, "Distace-based radio resource allocation for device to device communications," in *Proc. of IEEE 85th Vehicular Technology Conference (VTC Spring)*, pp. 1-5, 2017. Article (CrossRef Link).

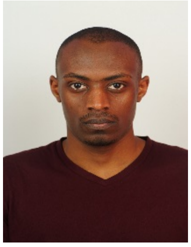
[15] P. Basuchowdhuri, "Greedy methods for approximate graph matching with applications for social network analysis," *LSU Master's Thesis 3105, 2009*. Article (CrossRef Link).

[16] N. Wormald, "Models of random regular graphs," *Surveys in combinatorics*, *London Math. Soc. Lecture Note Ser. 267*, *Cambridge Univ. Press*, pp.239-298, 1999. Article (CrossRef Link).

[17] J. Lofberg, “Automatic robust convex programming,” *Optimization methods and software*, vol. 27, no.1 pp 115-129, 2012. Article (CrossRef Link).

Guolin Sun received his B.S., M.S. and Ph.D. degrees all in Communication and Information System from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2000, 2003 and 2005 respectively. After Ph.D. graduation in 2005, Dr. Guolin has got eight years industrial work experiences on wireless research and development for LTE, Wi-Fi, Internet of Things (ZIGBEE and RFID, etc.), Cognitive radio, Localization and navigation. Before he joined the School of Computer Science and Engineering, University of Electronic Science and Technology of China, as an Associate Professor on Aug. 2012, he worked in Huawei Technologies Sweden. Dr. Guolin Sun has filed over 30 patents, and published over 30 scientific conference and journal papers, acts as TPC member of conferences. Currently, he serves as a vice-chair of the 5G oriented cognitive radio SIG of the IEEE (Technical Committee on Cognitive Networks (TCCN) of the IEEE Communication Society. His general research interest is software defined networks, network function virtualization, radio resource management.

Timothy Dingana received his Bachelor in Integrated Development Studies from University for Development Studies, Northern Region, Ghana, West Africa, in 2009. He also acquired MCSE and CCNA from IPMC Ghana, in 2010. He attained a Post-Graduate Diploma in Computer Science in JUALOM TECHNOLOGY INSTITUTE, Accra Ghana -2013. He graduated with MEng. in Computer Science and Technology from the University of Electronic Science and Technology of China (UESTC), Chengdu Sichuan, China (2018). From 2009 to 2011, he worked as a network support engineer for the Ghana Armed Forces General Headquarters Logistics, Accra Ghana. He is also a member of the Mobile Cloud-Net Research Team – UESTC. His interests include; Machine Learning, Mobile/Cloud Computing, Massive Datasets, LTE, 5G and SDN.



Sebakara Samuel Rene Adolphe received his Bachelor in Information and Communication Technology from University of Rwanda (former Umutara Polytechnique) Nyagatare-Rwanda, East Africa, in 2012. He received his MSc. Computer Science and Technology from the University of Electronic Science and Technology of China (UESTC) in 2017. He is currently studying his PhD in Computer Science and Technology at the University of Electronic Science and Technology of China (UESTC). From 2013, he worked as an Assistant Lecturer and Deputy Head of IT Department in Integrated Polytechnic (IPRC East) Ngoma-Rwanda. He is also a member of the Mobile Cloud-Net Research Team – UESTC. His interests include generally NFV, SDN, Mobile communication, Cloud Computing.

Gordon Owusu Boateng received his Bachelor in Telecommunications Engineering from the Kwame Nkrumah University of Science and Technology, Kumasi-Ghana, West Africa, in 2014. He is currently studying MSc. Computer Science and Technology in the University of Electronic Science and Technology of China (UESTC). From 2014 to 2016, he worked under sub-contracts for Ericsson (Ghana) and TIGO (Ghana). He is also a member of the Mobile Cloud-Net Research Team – UESTC. His interests include Mobile/Cloud Computing, 5G Wireless Networks, Data Mining, D2D Communications and SDN.