electronics

MDPI

Article

QoS-Aware Downlink Traffic Scheduling for Cellular Networks with Dual Connectivity

Haoru Su [1], Meng-Shiuan Pan [2,*] and Hung-Wei Mai [3]

[1] Department of Software Engineering, Faculty of Information Technology, Beijing University of Technology, Beijing 100021, China
[2] Department of Electronic Engineering, National Taipei University of Technology, Taipei 10608, Taiwan
[3] Department of Computer Science and Information Engineering, Tamkang University, New Taipei City 251301, Taiwan
* Correspondence: mspan@ntut.edu.tw

Abstract: In a cellular network, how to preserve users' quality of service (QoS) demands is an important issue. To provide better data services, researchers and industry have discussed the deployment of small cells in cellular networks to support dual connectivity enhancement for user equipments (UEs). By such an enhancement, a base station can dispatch downlink data to its surrounding small cells, and UEs that are located in the overlapping areas of the base station and small cells can receive downlink data from both sides simultaneously. We observe that previous works do not jointly consider QoS requirements and system capabilities when making scheduling decisions. Therefore, in this work, we design a QoS traffic scheduling scheme for dual connectivity networks. The designed scheme contains two parts. First, we propose a data dispatching decision scheme for the base station to decide how much data should be dispatched to small cells. When making a dispatching decision, the proposed scheme aims to maximize throughput and ensure that data flows can be processed in time. Second, we design a radio resource scheduling method, which aims to reduce dropping ratios of high-priority QoS data flows, while avoiding wasting radio resources. In this work, we verify our design using simulation programs. The experimental results show that compared to the existing methods, the proposed scheme can effectively increase system throughput and decrease packet drop ratios.

Keywords: cellular networks; dual connectivity; small cell; scheduling; quality of service (QoS)

Citation: Su, H.; Pan, M.-S.; Mai, H.-W. QoS-Aware Downlink Traffic Scheduling for Cellular Networks with Dual Connectivity. *Electronics* **2022**, *11*, 3085. https://doi.org/10.3390/electronics11193085

Academic Editors: Alexandros-Apostolos Boulogeorgos, Panagiotis Sarigiannidis, Thomas Lagkas, Vasileios Argyriou and Pantelis Angelidis

Received: 30 August 2022
Accepted: 22 September 2022
Published: 27 September 2022

1. Introduction

The popularity of wireless networks facilitates the development of various mobile applications, e.g., e-health care [1] and augmented reality (AR)/virtual reality (VR) [2]. People are used to enjoying these mobile services in their daily lives, and thus, the demands on the capacity and quality of mobile networks are still increasing. To provide broadband downlink data services, most countries have constructed 4G/5G cellular network systems specified by the third Generation Partnership Project (3GPP). In a 4G/5G cellular network, each *user equipment* (*UE*) will connect to a *master evolved Node B* (*MeNB*), i.e., the base station. The MeNB is responsible for assigning radio resources to UEs that connect to it. Furthermore, researchers and industry have discussed the deployment of small cells in cellular networks. Small cells are taken as *secondary evolved Node B* (*SeNBs*), which can provide dual connectivity enhancements for UEs. By such an enhancement, a UE that is located in the overlapping area of the MeNB and a SeNB can obtain radio resources from both sides simultaneously. In other words, UEs can receive more downlink data from both MeNB and SeNB, and thus the network capacity can be increased.

Figure 1 shows the protocol stack of the cellular network with dual connectivity enhancement defined in 3GPP specification 36.300 [3]. In this architecture, layer 1 is the physical layer (PHY). Layer 2 contains a media access control sublayer (MAC), radio link

control sublayer (RLC), and packet data convergence protocol sublayer (PDCP). The MeNB and SeNB are connected through an X2 interface, which may be ADSL or fiber connections. By this architecture, the MeNB decides how to split downlink data for UEs that connect to both the MeNB and SeNB. More specifically, for a dual-connectivity UE, the MeNB's PDCP layer decides how much data will be dispatched to the MeNB's RLC and to the corresponding SeNB's RLC layer, respectively. As shown in Figure 1, the separated downlink data will then be aggregated by the UE's PDCP layer. It is not hard to see how the MeNB dispatches downlink data will affect network performance. For example, assume that a UE u has connected to both the MeNB and a SeNB, and the observed signal quality on the SeNB is poor. In this situation, if the MeNB dispatches more downlink data to the SeNB, the SeNB needs to spend more radio resources to process the UE u's data, and thus, the system throughput will decrease. Some previous works (e.g., [4,5]) discussed data dispatching mechanisms for the MeNB. However, we found that the previous schemes do not consider SeNBs' capabilities and statuses when making dispatching decisions, and thus, the packet drop ratio increases because data dispatches to SeNBs may not be processed in time.

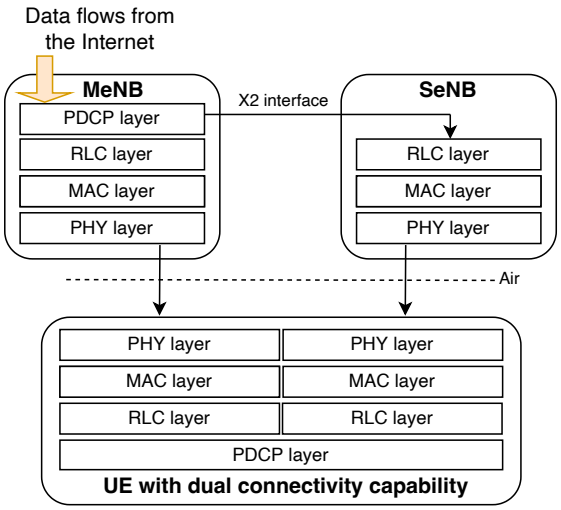

Figure 1. The protocol stack of a cellular network with dual connectivity enhancement.

In a 3GPP cellular network, the MAC layer of eNBs (i.e., the MeNB or SeNBs) assigns *radio resource blocks* (*RBs*) to disseminate downlink data for UEs every *transmission time interval* (*TTI*). Before assignment, the MAC layer collects channel quality indicator (CQI) values from UEs. Then, the MAC layer performs the adaptive modulation and coding algorithm (AMC) implemented by the network operator to determine the modulation and coding scheme (MCS) for UEs. Based on the MCS result, for a UE, the MAC layer can know the amount of data (in bits) that can be carried in an RB to the UE. Then, by referring to the remaining data amount of the UE, the MAC layer can determine how many RBs will be assigned to the UE in the next TTI. Obviously, RB assignment policies will affect network performance. In the literature, there are many QoS-aware MAC scheduling works (e.g., [6–9]), but most of them simply favor data flows with higher priorities and do not consider the status of buffered packets; as a result, packet drop ratio may be increased, and thus, overall system throughput cannot be increased.

As discussed above, downlink traffic scheduling in a cellular network with dual connectivity enhancement should consider (i) data dispatching between the MeNB and SeNBs and (ii) RB assignments in the MeNB and SeNBs. In this work, we design a QoS-aware scheduling solution that contains two parts. First, we propose a *QoS downlink data dispatching* (*QDD*) method for the MeNB. In QDD, when the MeNB receives downlink data from the Internet, the MeNB considers eNBs' capabilities and flows' QoS parameters to decide whether to dispatch flows to the corresponding SeNBs. The dispatching decisions aim to guarantee that separated downlink data can be processed in time at the SeNB. In QDD,

we model a linear programming formulation whose goal is to maximize network throughput, and dispatching decisions can be obtained by solving the proposed formulation. Second, we propose a *QoS eNB scheduling* (*QED*) method for the MeNB and SeNBs. To reduce packet loss, the QED attempts to give more opportunities to those high-priority packets that are going to expire. However, before giving resources, the QED evaluates channel qualities and packets' remaining transmissions to ensure those high-priority traffics can be processed before deadlines. In this work, we verify our design using simulation programs. The experimental results show that compared to the existing methods, the proposed scheme can effectively increase system throughput and decrease packet drop ratios. The contributions of this paper are summarized below.

- We propose QDD and QED methods to schedule downlink traffics for cellular networks with dual connectivity enhancement. The proposed schemes can satisfy QoS requirements while increasing system throughput and decreasing packet drop ratios.
- In previous works, the MeNB dispatches downlink data to SeNBs without considering SeNBs' capabilities; as a result, the data dispatched to SeNBs may not be processed in time. The proposed QDD method makes dispatching decisions by referring to SeNBs' capabilities and statuses. This design ensures that data flow will not arbitrarily forward to SeNBs.
- When allocating radio resources (i.e., RBs), most previous works favor high-priority data flows, and thus, resources may be wasted to disseminate packets that cannot be fully processed before timeout. The proposed QED method evaluates the status of buffered GBR packets before giving resources to high-priority data flows. This design can avoid wasting radio resources on packets that are going to expire soon.
- Based on our surveys, this is the first work to jointly consider QoS requirements and system capabilities when making data dispatching and RB assignment decisions. Furthermore, the proposed solutions can be compatible with the 3GPP standards.

The remainder of this paper is organized as follows. Section 2 reviews previous studies on traffic scheduling in dual connectivity networks and MAC scheduling of the MeNB and SeNBs. Section 3 describes network models and system flows of this work. Then, Section 4 and Section 5 present the proposed QDD and QED methods, respectively. Next, Section 6 evaluates the network performance of the proposed scheme and compares its performance with previous schemes. Finally, Section 7 concludes this paper and outlines the future work.

2. Related Works

2.1. Traffic Scheduling in Dual Connectivity Networks

In the following, we review previous works on cellular networks with dual connectivity capability. The study [10] introduces a framework to facilitate connectivity mode selections for UEs. By the proposed scheme, each UE decides whether to enable the dual connectivity mode based on its energy efficiency. The proposed scheme can reduce congestion and improve the QoS of UEs, but the proposed scheme only focuses on uplink transmissions. In reference [11], the authors use non-orthogonal multiple access (NOMA) technology to improve the energy efficiency of MeNB and SeNB. The authors define a linear programming formulation, whose goal is to minimize power consumption when disseminating downlink data. However, the formulation cannot be used to make traffic dispatching decisions. The authors in [12] design a data splitting scheme for heterogeneous networks. The designed scheme utilizes the concept of water-filling to allocate radio resources to dual-connected UEs. By the scheme, the throughput of UEs can be effectively increased, but the network architecture is not compatible with 3GPP cellular networks. The constrained Markov decision process to make the traffic splitting decision is discussed in [13]. The designed scheme aims to minimize packet delays. However, UEs' capabilities and signal qualities are not considered when modeling the network, and thus, delays cannot be precisely evaluated. Furthermore, in reference [14], the authors define the gain of the network when UEs receive downlink data from the MeNB and a SeNB simultaneously.

Then, based on the estimated signal qualities of UEs and delays of data flows, the proposed scheme maximizes network gain by deciding how to dispatch data to SeNBs. In [15], a downlink data dispatching method is introduced, whose goal is to save the power of the backhaul links. In the proposed scheme, the MeNB dispatches data to SeNBs if energy consumption of the network or receiving qualities of UEs can be improved. In reference [4], each dual-connected UE reports its receiving status to the MeNB every TTI. After collecting UE reports, the MeNB uses predefined functions to make traffic splitting decisions. In reference [5], each UE first decides which SeNB it wants to connect to. After confirming the UE's choice, the MeNB dispatches data to the corresponding SeNB according to the network load and fairness. Furthermore, in [16], the authors present a data dispatching method for LTE-A networks with dual connectivity enhancement. In the proposed scheme, network parameters (such as remaining data volumes and connection statuses) are used to model a linear programming formulation. The MeNB refers to solutions of the linear formulation to dispatch data to SeNBs. However, the proposed formulation does not take into account SeNB's capability. Next, the authors in [17,18] solve the TCP out-of-order phenomenon in dual connectivity networks. In [18], two splitting algorithms are designed to reduce the probability of out-of-order receptions on the UE sides, but UEs have to report frequently. In [17], the authors utilize the Raptor code concept to resolve an out-of-order problem. By Raptor codes, UEs can decode source data after receiving enough encoded data. However, the proposed scheme cannot be used to support networks with dense traffic loads.

The above previous works propose traffic scheduling methods in dual connectivity networks. Table 1 summarizes the comparisons of previous works with the proposed QDD in three aspects. Two previous works (i.e., [10,13]) are not designed to make downlink traffic dispatching decisions. Although references [12,17,18] describe dispatching methods, these schemes do not take QoS parameters into consideration. Furthermore, the other works do not consider the SeNB's capability when making dispatching decisions; as a result, downlink data (that sent to SeNBs) may not be processed in time. From the above analysis, we can see that the proposed QDD offers the most complete solutions for the target network.

Table 1. Comparison of prior works with QDD.

Reference	Perform Dispatch	Consider QoS	Consider SeNB Capability
Ref. [4]	✓	✓	
Ref. [5]	✓	✓	
Ref. [10]		✓	
Ref. [17]	✓		
Ref. [16]	✓	✓	
Ref. [15]	✓	✓	
Ref. [12]	✓		
Ref. [18]	✓		
Ref. [13]	✓	✓	
Ref. [14]	✓	✓	
Ref. [11]		✓	
QDD	✓	✓	✓

2.2. MAC Scheduling for eNBs

In a 3GPP cellular network, the eNB's MAC layer determines RB allocations every TTI, and there are many QoS MAC scheduling methods in the literature. The authors in [8]

propose a QoS scheduling method to allow data flows with higher priorities and better signal quality to earn additional radio resources. In [19], the proposed scheme divides data flows into two categories. The category that contains real-time traffic will be scheduled earlier. The schemes in [8,19] can increase throughput of QoS data, but do not consider the remaining transmission time of packets when scheduling. In [20], the proposed scheme sorts head-of-line delays of downlink packets in ascending order and then decides RB assignments to the corresponding data flows based on the sorted result. The authors in [9] discuss QoS scheduling in small cell networks. The proposed method dynamically adjusts UEs' priorities to ensure that high-priority data can obtain more resources. In [21], a two-stage scheduling method is introduced. The first stage schedules based on packets' remaining transmission times, and the second stage schedules low-priority data flows based on the proportional fair (PF) concept. In [7], the authors propose a two-level scheduling structure. The first level applies game theory to achieve fairness between data flows. The second level then utilizes the knapsack algorithm to assign radio resources. Moreover, in [22], the authors propose a linear programming formulation to make downlink scheduling decisions. The proposed formulation includes jitter, queuing delay, and priority as constraints, and the goal is to satisfy QoS demands of data flows. The authors in [23] consider the scheduling of real-time and non-real-time traffics. The proposed scheme can adapt to various constraints (e.g., head-of-line delay, queue length, channel quality) and try to balance the transmissions between these two traffic types. In [24], the authors propose a scheduling method that aims to minimize packet delay violations. The proposed formulation considers packet arrival rates, packet loss ratio, and head-of-line delays when making scheduling decisions. However, the proposed scheme uses static settings to differentiate traffics. In reference [25], the authors propose a downlink scheduling method for the existence of device-to-device (D2D) links. The proposed scheme can help increase network throughput, but QoS requirements are not addressed. The authors in [6] consider downlink resource scheduling in the network scenario with hybrid beamforming support. The proposed heuristic algorithm can increase network throughput by finding suitable beams to disseminate downlink data. In the proposed scheme, packets that are going to expire can be scheduled earlier, but channel qualities and remaining transmission times are not considered when scheduling. The authors in reference [26] design a scheduling method that considers the QoS requirements and fairness between data flows. When scheduling, the proposed scheme estimates possible packet drops of GBR traffic flows, and then attempts to borrow resources allocated to other flows. However, it is possible that the other flows' resources cannot be borrowed, and thus, the GBR packets will still be dropped. In this work, when allocating resources to packets that are going to expire, the proposed scheme will evaluate the status of buffered data and then allocate resources. By this design, the proposed scheme can effectively decrease packet drop ratio.

The above previous works design MAC scheduling methods for the MeNB and SeNBs. Table 2 summarizes the comparisons of previous works with the proposed QED in three aspects. Again, the proposed QED offers the most complete solutions for eNBs. More specifically, when allocating radio resources, the proposed QED scheme considers (i) the network situation (i.e., channel quality), (ii) fairness between traffic flows, and (iii) the packets' remaining transmission times. We can see that the scheme designed in reference [26] has features similar to ours, but it may not effectively allocate resources to GBR flows that are going to expire.

Table 2. Comparison of prior works with QED.

Reference	Consider QoS	Consider Fairness	Consider Packet Status
Ref. [22]	✓	✓	
Ref. [6]			✓
Ref. [7]	✓		

Table 2. *Cont.*

Reference	Consider QoS	Consider Fairness	Consider Packet Status
Ref. [8]	✓		
Ref. [9]	✓		
Ref. [25]		✓	
Ref. [21]			✓
Ref. [23]	✓	✓	
Ref. [20]			✓
Ref. [24]	✓		
Ref. [19]	✓		
Ref. [26]	✓	✓	✓
QED	✓	✓	✓

3. Network Model and System Flow

Figure 2 shows the network scenario. The network is composed of one MeNB and l SeNBs. These l SeNBs are modeled as the set $S = \{s_1, s_2, \ldots, s_l\}$. The operating frequencies of the MeNB and l SeNBs are different, and thus, the wireless signals between the MeNB and a SeNB will not interfere with each other. The MeNB and l SeNBs are connected by X2 interfaces. The link capacities of these interfaces are $R_{bh}(s_1), R_{bh}(s_2), \ldots, R_{bh}(s_l)$, respectively. In the network, there are m UEs, which are modeled by the set $U = \{u_1, u_2, \ldots, u_m\}$. The set U can be divided further into U^M and U^C sets. A UE located in the U^M set indicates that the UE only connects to the MeNB. On the other hand, a UE located in the U^C set means that the UE connects both to the MeNB and a SeNB at the same time. For example, in Figure 2, there are two SeNBs, i.e., $l = 2$, and five UEs, i.e., $m = 5$. Based on the above model, the sets $U^M = \{u_2, u_5\}$ and $U^C = \{u_1, u_3, u_4\}$.

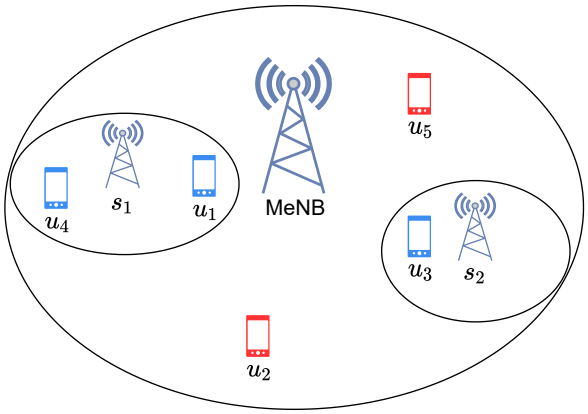

Figure 2. The network scenario.

According to the 3GPP specifications, each data flow will be carried by a radio bearer. A UE can initiate multiple radio bearers, i.e., a UE can have multiple data flows at the same time. However, to facilitate presentation, we simply assume that those m UEs in the network will initiate m radio bearers accordingly. These bearers are represented by a set $B = \{b_1, b_2, .., b_m\}$, where the bearer b_i belongs to the UE u_i. The set B can also be divided into B^M and B^C sets. For a bearer b_i, its incoming data rate is $r_{in}(b_i)$. Furthermore, the bearer b_i will be assigned to a 3GPP QoS class identifier (QCI) value [27]. The QCI information contains the bearer's priority level, delay constraint, and packet drop ratio requirements. We model the delay and packet drop ratio constraints of b_i as $\mathcal{Q}_d(b_i)$ and $\mathcal{Q}_l(b_i)$, respectively. When a packet of b_i arrives to the MeNB's PDCP layer at time t,

this packet has to be sent to the corresponding UE no later than time $t + \mathcal{Q}_d(b_i)$. Otherwise, this packet will be dropped. Besides, based on the QCI value, traffic can further be divided into the guaranteed bit rate (GBR) and non-guaranteed bit rate (non-GBR). The GBR traffics are delay sensitive, which are used to support real-time traffics (e.g., video stream and online game). On the other hand, the non-GBR traffics can tolerate more delay and the data size may be larger. According to the 3GPP QCI settings, GBR traffic types will have higher priorities than non-GBR ones. Therefore, in this work, GBR traffics will be served earlier.

Figure 3 indicates the proposed framework. As mentioned in Section 1, the MeNB's PDCP layer will receive UEs' downlink packets from the Internet. The MeNB buffers the received data and then makes dispatch decisions using the proposed QDD method (in Section 4). In our design, the MeNB periodically executes QDD every $I_t = k \times TTI$ ms, where k is a predefined system parameter. By QDD, part of the received packets for these dual-connectivity UEs will be sent to the corresponding SeNBs. Furthermore, the downlink data will be sent to the MAC layer of the MeNB and SeNBs. Then, the MAC layer adopts the proposed QED method to allocate its radio resource blocks (RBs) to UEs underneath the corresponding MeNB or SeNB every TTI.

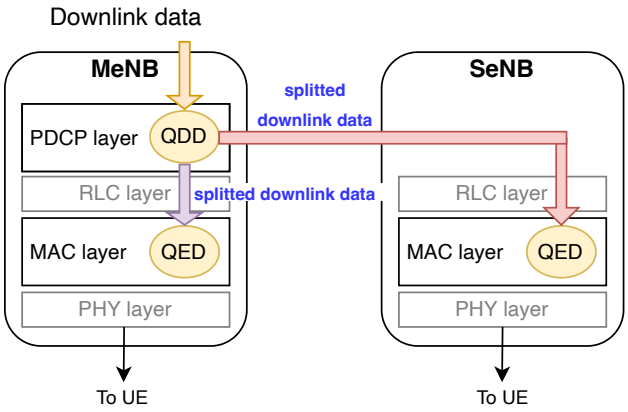

Figure 3. The proposed framework.

4. Proposed QoS Downlink Data Dispatching (QDD) Method

In this section, we introduce the proposed QDD method to dispatch downlink data flows between the MeNB and SeNBs. The QDD method is composed of a linear programming formulation, in which the goal is to maximize the network throughput under the considerations of network capabilities and QoS requirements. Recall that for a bearer b_j, its incoming data rate is $r_{in}(b_j)$. In QDD, if a bearer b_j belongs to B^C, the QDD will determine a $r_{sp}(b_j)$ value, which represents the data splitting rate. More specifically, after determining $r_{sp}(b_j)$, the MeNB will dispatch data flow with a rate $r_{sp}(b_j)$ to the corresponding SeNB, and the MeNB itself will handle the remaining data with a rate $(r_{in}(b_j) - r_{sp}(b_j))$. Before showing the detailed formulation of QDD, we define the following parameters.

1. RB_{max} and $RB_{max}(s_k)$: The total number of RBs that can be used by the MeNB and SeNB s_k during I_t, respectively.
2. $M_{u2s}(u_i) = s_k$: This relationship represents that the UE u_i is connected to the SeNB s_k.
3. $D_r^M(u_i)$ and $D_r^S(u_i)$: The amount of UE u_i's downlink data currently stored in the MeNB and the corresponding SeNB, respectively.
4. $O_m(u_i)$ and $O_s(u_i)$: The amount of data that can be carried in an RB for UE u_i allocated by the MeNB and the corresponding SeNB, respectively. These parameters can be estimated by averaging the CQI reports of u_i in the previous period. Further, let rs be the size of an RB in units of seconds. From $O_s(u_i)$, we can further estimate transmission capabilities $\hat{O}_s(u_i)$ (in bps) of u_i, where $\hat{O}_s(u_i) = \frac{O_s(u_i)}{rs}$.
5. $Rb^M(u_i)$ and $Rb^S(u_i)$: The number of RBs that the MeNB and the corresponding SeNB expect to assign to UE u_i, respectively. Note that these two parameters are unknown

variables in our formulation. In other words, these two variables can be derived after solving the following linear programming formulation.

6. $\mathcal{D}_p^M(u_i)$ and $\mathcal{D}_p^S(u_i)$: The amount of UE u_i's data that can be consumed by the MeNB and the corresponding SeNB in the next I_t, respectively. More specifically, $\mathcal{D}_p^M(u_i)$ and $\mathcal{D}_p^S(u_i)$ can be obtained by the following two equations:

$$\mathcal{D}_p^M(u_i) = O_m(u_i) \times Rb^M(u_i)$$

$$\mathcal{D}_p^S(u_i) = O_s(u_i) \times Rb^S(u_i)$$

7. $R_{eff}^M(u_i)$ and $R_{eff}^S(u_i)$: UE u_i's effective input data rates through the MeNB and the corresponding SeNB, respectively. These two parameters can be divided into the following two cases.
   - If $u_i \in U^M$, i.e., u_i is only served by the MeNB, $R_{eff}^M(u_i) = r_{in}(b_i)$ and $R_{eff}^S(u_i) = 0$.
   - If $u_i \in U^C$, i.e., u_i is served by the MeNB and the SeNB $M_{u2s}(u_i)$, $R_{eff}^M(u_i) = r_{in}(b_i) - R_{eff}^S(u_i)$, and $R_{eff}^S(u_i) = r_{sp}(b_i)$.

In QDD, the objective of the proposed linear programming formulation is to maximize the network throughput, i.e., the total amount of data allocated to all UEs connected to the MeNB and SeNBs in the next I_t interval. The proposed formulation is as follows.

$$\max \sum_{u_i \in U} \mathcal{D}_p^M(u_i) + \mathcal{D}_p^S(u_i) \tag{1}$$

s.t.

$$C1: r_{sp}(b_i) \le r_{in}(b_i), \forall b_i \in B^C$$
$$r_{sp}(b_i) \le 0, \forall b_i \in B^M$$
$$C2: \mathcal{D}_p^M(u_i) \le R_{eff}^M(u_i) \times I_t + D_r^M(u_i), \forall u_i \in U$$
$$C3: \mathcal{D}_p^S(u_i) \le R_{eff}^S(u_i) \times I_t + D_r^S(u_i), \forall u_i \in U^C$$
$$C4: \sum_{\substack{u_i \in U^C \\ b_i \in B^C}} \{r_{sp}(b_i) | M_{u2s}(u_i) = s_k\} \le R_{bh}(s_k), \forall s_k \in S$$
$$C5: \sum_{u_i \forall U} Rb^M(u_i) \le Rb_{\max}, \text{ for the MeNB}$$
$$C6: \sum_{u_i \forall U} \{Rb^S(u_i) | M_{u2s}(u_i) = s_k\} \le Rb_{\max}(s_k), \forall s_k \in S$$
$$C7: \mathcal{Q}_d(b_i) \ge \frac{r_{sp}(b_i) \times I_t}{R_{bh}(M_{u2s}(u_i))} + \frac{r_{sp}(b_i) \times I_t + D_r^S(u_i)}{\hat{O}_s(u_i)}, \forall b_i \in B^C \cap \forall u_i \in U^C$$

In the following, we discuss the constraints of the above formulation. First, in constraint C1, if the bearer $b_i \in B^C$, this means that UE u_i connects to the MeNB and a SeNB simultaneously, and thus, the data flow of b_i can be split. Therefore, $r_{sp}(b_i)$ must be less than or equal to the b_i's input data rate $r_{in}(b_i)$. On the other hand, if $b_i \in B^M$, it means that UE u_i only connects to the MeNB, and thus, $r_{sp}(b_i) \le 0$. Second, in constraint C2, for any UE $u_i \in U$, the amount of downlink data $\mathcal{D}_p^M(u_i)$ that can be consumed by the MeNB in the next I_t interval must be less than or equal to the sum of (i) the amount of data that can be transmitted by the MeNB in the next I_t interval, i.e., $R_{eff}^M(u_i) \times I_t$, and (ii) the amount of u_i's remaining data amount $D_r^M(u_i)$ in the MeNB. Third, in constraint C3, for the UE $u_i \in U^C$ (i.e., u_i connects both to the MeNB and a SeNB at the same time), the amount of downlink data $\mathcal{D}_p^S(u_i)$ that can be consumed by the SeNB in the next I_t interval must be less than or equal to the sum of (i) the amount of data that can be transmitted by the SeNB in the next I_t interval, i.e., $R_{eff}^S(u_i) \times I_t$, and (ii) the amount of u_i's remaining data

amount $D_r^S(u_i)$ in the SeNB. Fourth, in constraint C4, for a SeNB s_k, the total data rates that split from the MeNB to s_k should be less than or equal to the X2 link capacity $R_{bh}(s_k)$. Fifth, constraints C5 and C6 specify that the number of RBs allocated to UEs that connect to the MeNB and the corresponding SeNB s_k should not be larger than $Rb_{\max}$ and $Rb_{\max}(s_k)$, respectively. Sixth, constraint C7 requires that if the flow of the bearer b_i' is split to the SeNB, the corresponding packets handled by the SeNB should be delivered to the UE u_i before the delay constraint $\mathcal{Q}_d(b_i)$. When dispatching data to the SeNB, the possible delay includes: (i) The transmission delay caused by the X2 interface, which is the total amount of data dispatched to SeNB in the next I_t interval $r_{sp}(b_i) \times I_t$ divided by the corresponding X2 interface link capacity $R_{bh}(M_{u2s}(u_i))$. (ii) The processing time of the SeNB. For a UE $u_i \in U^C$, the SeNB needs to process u_i's remaining data $D_r^S(u_i)$ and the upcoming data $r_{sp}(b_i) \times I_t$ in the next I_t interval. The processing time can be estimated by the total amount of data divided by the u_i's estimated transmission capability $\hat{O}_s(u_i)$.

According to the above formulation, unknown variables include $Rb^M(u_i)$ and $Rb^S(u_i)$, $\forall u_i \in U$, and $r_{sp}(b_i)$, $\forall b_i \in B$. We can see that the number of unknown variables is less than the total number of equations. Thus, we can find an optimal solution using a linear programming solver. In every I_t, the MeNB fills in the known variables of the above equation and derives the results of $r_{sp}(b_i)$, $\forall b_i \in B$. Then, the MeNB can follow the results to dispatch downlink data to the corresponding SeNBs in the upcoming I_t. Note that in the above formulation, a UE u_i only associates with a bearer b_i. As we mentioned above, a UE may initiate multiple data flows at the same time. Therefore, it is not hard to see that our formulation can adapt to this scenario by adding virtual UEs.

5. Proposed QoS eNB Scheduling (QED) Method

In this section, we introduce the proposed MAC scheduling method, named QED, which aims to prevent GBR packets to be dropped. The basic idea of QED is that if a GBR packet satisfies the designed criteria (described later), this packet will be moved to a *guarantee to transmit* (*GTT*) set. For those packets in the GTT set, the QED will decide their transmit sequences. Then, when assigning RBs, the QED first allocates RBs to serve packets in GTT. After processing packets in GTT, if there are remaining RBs, the QED allocates RBs to the other data flows in the system.

In the following, we describe the designed QED method in detail. In every TTI, the QED first discards the expired packets and then removes them from the system. The QED then checks the existing GBR data flows. For a GBR data flow, the QED first selects the packets whose remaining transmission times are less than $\hat{T}_n$ TTIs. For each of the selected packets, the QED performs the following operations: (Assume that the QED is now processing the packet p, which belongs to UE u_i and the corresponding bearer b_i.) First, the QED calculates p's needed transmission time and remaining transmission time as follows.

- p's needed transmission time $\mathcal{E}(p)$: Let the packet length of p be L_p and $\hat{O}(u)$ be the estimated transmission capability (defined in Section 4). The QED estimates p's needed transmission time as:
$$\mathcal{E}(p) = \frac{L_p}{\hat{O}(u_i)}.$$
- p's remaining transmission time $\mathcal{R}(p)$: Let $\mathcal{W}(p)$ be the p's waiting time in the system. Recall that $\mathcal{Q}_d(b_i)$ is the corresponding delay constraint of the bearer b_i (defined in Section 3). The QED calculates p's remaining transmission time as:
$$\mathcal{R}(p) = \mathcal{Q}_d(b_i) - \mathcal{W}(p).$$
Note that $\mathcal{W}(p)$ can be derived by subtracting the current time with the time instant that the packet p arrived at the PDCP layer of the MeNB.

Based on the definitions of $\mathcal{E}(p)$ and $\mathcal{R}(p)$, if $\mathcal{E}(p) < \mathcal{R}(p)$, this implies that the packet p cannot be processed on time and p will not be scheduled. Otherwise, the packet p will be put into the GTT set if the packet p satisfies the following inequality:

$$\frac{\mathcal{R}(p)}{\mathcal{Q}_d(b_i)} \times \frac{1}{-\log \mathcal{Q}_l(b_i)} \leq \frac{\mathcal{E}(p)}{\mathcal{R}(p)} \times \frac{1}{-\log \mathcal{L}(b_i)} \tag{2}$$

In Equation (2), the $\mathcal{L}(b_i)$ represents the measured packet drop ratio of bearer b_i, and $\mathcal{Q}_l(b_i)$ represents the allowable packet drop ratio of b_i (defined in Section 3). Note that $\mathcal{L}(b_i)$ may change over time. Since loss or drop ratios may be small, we use the logarithm to force $\mathcal{Q}_l(b_i)$ and $\mathcal{L}(b_i)$ to be positive values. Next, we explain the rationale of Equation (2). The design of Equation (2) is a dynamic threshold for the packet p. On the left side of Equation (2), the $\mathcal{Q}_d(b_i)$ and $\mathcal{Q}_l(b_i)$ are fixed values. The left side of Equation (2) will be smaller if the remaining transmission of p, i.e., $\mathcal{R}(p)$ becomes smaller, and the packet p will have more chances to enter GTT. Besides, if the $\mathcal{Q}_l(b_i)$ is smaller, the loss ratio requirement is more critical, and thus, the packet p may enter GTT more easily. Moreover, on the right side of Equation (2), if $\mathcal{R}(p)$ becomes smaller, this equation allows the packet p to have a higher probability to enter GTT. When the needed transmission time $\mathcal{E}(p)$ is higher, there may be two possibilities: (i) the size of p is larger or (ii) the signal quality of destination UE becomes weak. Thus, when $\mathcal{E}(p)$ is higher, the packet p can have more chances to enter GTT. Finally, if the measured drop ratio $\mathcal{L}(b_i)$ is higher, the packet p is allowed to enter the GTT more easily.

When a packet p is in the GTT set, the QED calculates a priority value $\hat{Q}(p)$ for p using the following equation:

$$\hat{Q}(p) = \frac{qci(b_i) \times \mathcal{R}(p)}{cqi(u_i)} \tag{3}$$

In Equation (3), $qci(b_i)$ is the QoS QCI value of the corresponding bearer b_i, and $cqi(u_i)$ is the CQI value reported by the destination UE u_i. In our design, if the $\hat{Q}(p)$ is smaller, the packet p can have a higher priority. Next, we discuss the design of Equation (3). According to the 3GPP standard, when the QCI value is smaller, the priority of the corresponding data flow is higher, and thus, the corresponding packet p can have higher priority. Then, when the remaining transmission time $\mathcal{R}(p)$ decreases, the packet p can have higher priority. Finally, if the reported CQI $cqi(u_i)$ is higher, this means that the corresponding UE u_i's signal quality is better, and this packet can have a higher priority.

After calculating the priority value for every packet in GTT, the QED sorts packets in GTT according to their priorities in nondecreasing order. In every TTI, when scheduling RBs, the packets in GTT will be assigned to RBs earlier. The QED then allocates the remaining RBs to existing data flows using the proportional fair (PF) scheduling method [28]. Furthermore, Algorithm 1 outlines the overall procedures of the QED algorithm.

Algorithm 1 The QED algorithm

```
Discard those packets that have expired;
foreach GBR flow f in the system do
    foreach packet p in GBR flow f do
        if Remaining transmission time of p is less than T̂_n and E(p) ≥ R(p) then
            p will be put to GTT set;
Sort packets in GTT set according to their priority;
while there are remaining RBs do
    Assign enough RBs to disseminate first packet in GTT set;
    Remove the first packet in GTT set;
if there are remaining RBs then
    Assign RBs to remaining flows by proportional fair (PF) algorithm;
```

6. Simulation Results

In this work, we evaluate our designs using simulation programs implemented by C programming language. In our simulations, the network area is a circle with radius 2 km, and the MeNB is located in the center. There are l SeNBs in the network, and the distance between MeNB and SeNB is 1 km. In every TTI, the MeNB and SeNB can assign 100 and 30 RBs to UEs, respectively. All UEs in the network have dual connectivity capabilities. The initial locations of UEs are randomly distributed in the network area, and each UE will move to a random direction every 0.5 s with a default moving speed of 5 km/h. Based on the locations of the UEs, some UEs may only be connected to the MeNB. Each UE has three downlink traffics: voice, video, and constant bit rate (CBR) with rates of 8.4, 242, and 12 kbps, respectively. The voice traffic and video traffic are taken as GBR types in our simulations, and voice traffic has higher priorities than video traffic. Furthermore, our simulator adopts the urban macro-cell model as the propagation model, where the path loss is calculated by $128.1 + 37.6 \log(L)$, the thermal noise is set to -174 dBm/Hz, and the fast fading model is Rayleigh. The maximum transmit power levels of the MeNB and SeNB are set to 46 and 23 dBm, respectively.

We compare the proposed QDD data dispatching scheme with (i) the DTS scheme in [16] and (ii) the fixed dispatching scheme (denoted by FIX). The FIX scheme always dispatches 50% of downlink data to the SeNB. We compare QDD with DTS because DTS uses similar concepts to dispatch downlink data to SeNBs. Moreover, we compare the proposed QED scheduling scheme with the DFS scheme in [26] since the DFS has similar features with QED as mentioned in Section 2.2. The DFS jointly considers QoS and maximizing throughput while scheduling. We aim to observe the advantage of QED, which considers the status of GBR packets before giving resources to those packets that are going to expire. In our simulations, we show the results on different combinations of the above data dispatching and MAC scheduling schemes. Furthermore, we solve the proposed linear programming formulation of QDD using the lpsolve solver [29]. The QDD makes dispatching decisions (i.e., to solve the proposed linear formulation) every $I_t = 100$ ms. Furthermore, for QED, we set $\hat{T}_n = 10$ ms, i.e., a GBR packet can enter QTT when its remaining transmission time is less than 10 ms. Each simulation result is derived by averaging the results produced within a 60 s simulation time.

Figure 4 shows the simulation results when the network has one MeNB and one SeNB, and we vary the number of UEs in the network. First, Figure 4a,b indicates the results of voice and video throughput, respectively. We can see that in Figure 4a,b, the proposed QDD+QED scheme can outperform the others. On average, the QDD+QED can be 9%, 14%, 36%, and 44% (resp., 5%, 15%, 28%, and 49%) better than DTS+QED, DTS+DFS, FIX+QED, and FIX+DFS in voice (resp. video) throughput, respectively. In DTS, the authors do not consider QoS parameters when making dispatching decision. As a result, DTS's GBR traffic throughput will be lower since more packets may be dropped in the SeNB. On the other hand, we can also see that GBR traffic throughput of DTS+QED can be better than that of DTS+DFS. This result indicates that the proposed QED scheme can indeed facilitate the scheduling of GBR traffics. The FIX+QED and FIX+DTS will have the worst GBR throughput. This implies that a well-designed dispatching scheme can facilitate the increase of GBR downlink data throughput. Furthermore, Figure 4c indicates the throughput of CBR traffics (i.e., non-GBR traffics). We can see that the CBR throughput of QDD+QED will be lower than that of DTS+DFS. This is because both DTS and DFS aim to maximize network throughput without considering the status of GBR traffics. Therefore, the combination of DTS+DFS may allocate more resources to non-GBR traffic. Finally, Figure 4d shows the overall system throughput. Since there are more video traffics in the simulated networks, the overall system throughput of the proposed scheme can be better than the others. On average, the QDD+QED can be 9%, 12%, 22%, and 32% better than DTS+QED, DTS+DFS, FIX+QED, and FIX+DFS in system throughput, respectively.

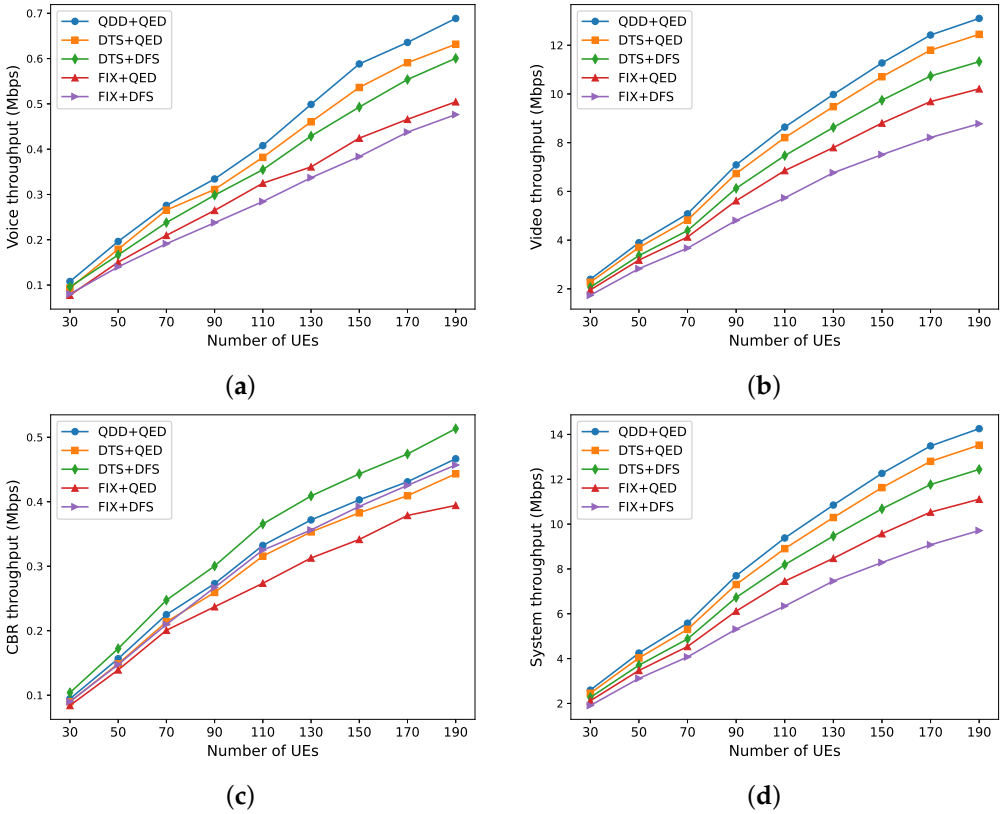

Figure 4. Simulation results on (**a**) voice, (**b**) video, (**c**) CBR, and (**d**) system throughput when varying number of UEs.

Figure 5a,b indicates the packet drop ratios of video and CBR traffics of Figure 4b and Figure 4c, respectively. In the simulation, when the number of UEs is less than 90, the drop ratios of video traffic and CBR traffic will be less than 0.5%, and thus, we do not show the results when the number of UEs is less than 90. From Figure 5a, QDD+QED will have the lowest packet drop ratios on video traffic, so the video throughput of QDD+QED can be higher (as discussed above). Again, from Figure 5b, DTD+DFS will have the lowest packet drop ratios on CBR traffic; thus, the CBR throughput of DTD+DFS can be higher (as discussed above). From these two results, we can see that, by the QDD+QED scheme, video traffic drop ratios can always be lower than CBR traffic drop ratios. This result demonstrates that the proposed scheme can indeed help facilitate the transmissions of GBR traffic. Moreover, Figure 5c indicates the system packet drop ratio. From the result, we can observe that the proposed QDD+QED can achieve the lowest drop ratio for the entire network. From the result, on average, the QDD+QED can be 6%, 3%, 23%, and 25% better than DTS+QED, DTS+DFS, FIX+QED, and FIX+DFS in system packet drop ratios, respectively.

Next, we fix the number of SeNBs and UEs to be 1 and 55, respectively, and we observe the results when varying the UEs' moving speeds. Figure 6 indicates the results. From Figure 6a,b, we see that the proposed QDD+QED can outperform the others in GBR throughput. We also see that when the moving speeds of UEs become faster, the GBR traffic throughput decreases accordingly. This is because when UEs move farther away from the SeNB, their downlink data (stored in the SeNB) may not be disseminated to the corresponding UEs in time. Since both QDD and QED consider traffic delay constraints, the throughput of GBR traffic can still be higher than the others. Similar to the above discussions, DTS+DFS can perform better than our scheme on CBR throughput (as shown in Figure 6c), but according to Figure 6d, the proposed scheme can still outperform other schemes on system throughput.

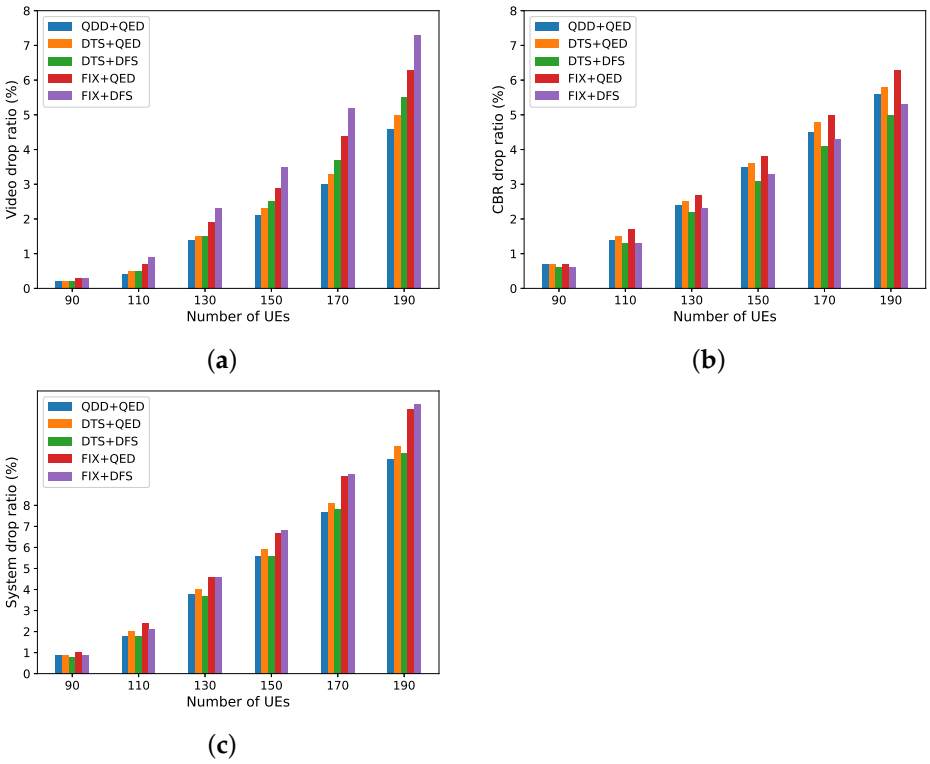

Figure 5. Simulation results on (**a**) video and (**b**) CBR (**c**) system packet drop ratios when varying number of UEs.

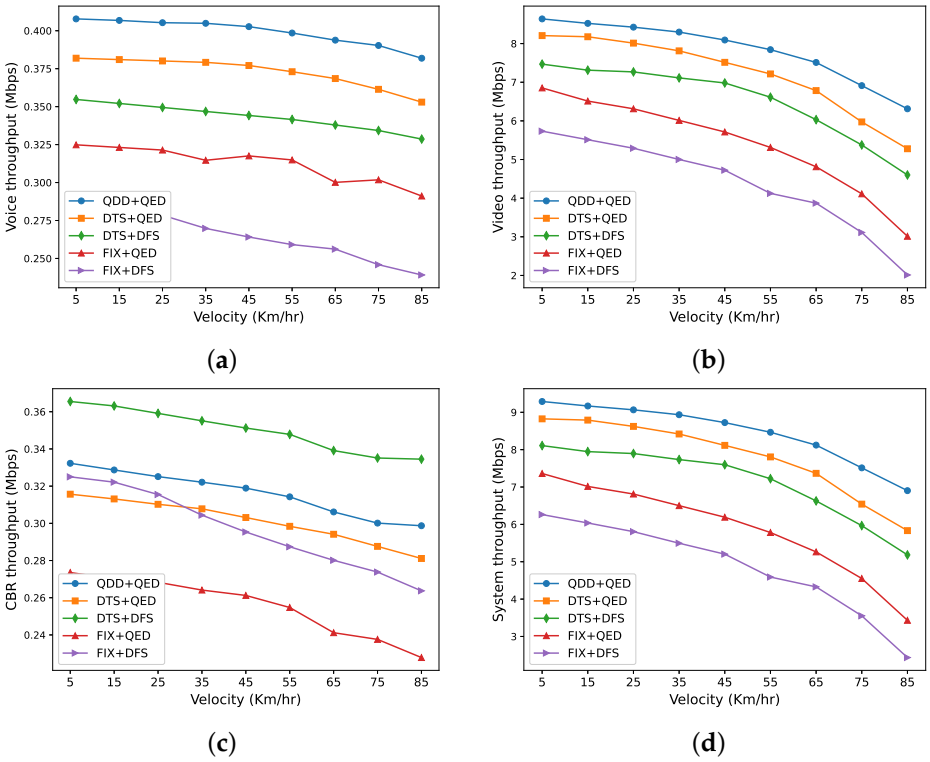

Figure 6. Simulation results on (**a**) voice, (**b**) video, (**c**) CBR, and (**d**) system throughput when varying velocities of UEs.

Finally, we simulate the scenario that the network has l SeNBs, where $l = 1 \ldots 4$, and in the network, there are $l \times 55$ UEs. Figure 7 shows the simulation results. When there are more SeNBs in this network, the throughput values will increase accord-

ingly due to the increased number of UEs. From the results, we can see that the proposed scheme can still outperform the others when traffics are GBR type. Similar to the above simulations results, the DTS+DFS can be slightly better than our scheme in CBR throughput. However, the proposed scheme can perform better than the others on system throughput. Furthermore, recall that the formulation of QDD considers the capability of SeNBs. These simulation results demonstrate that the proposed formulation can be adopted for the network with multiple SeNBs.

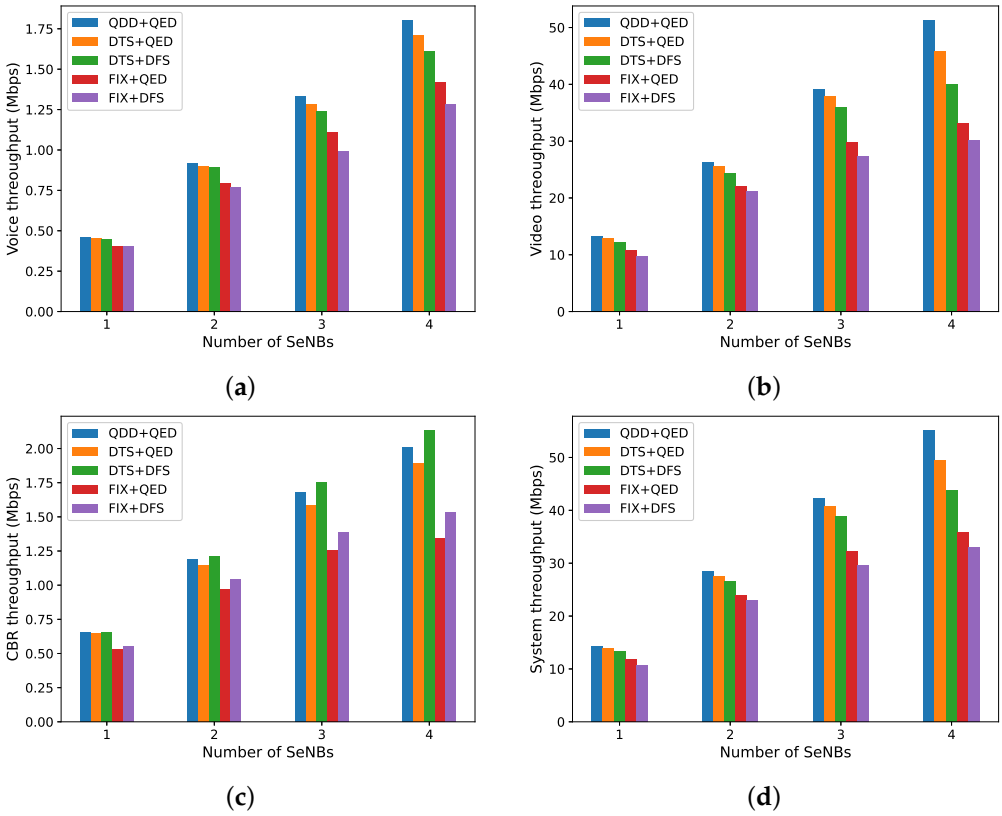

Figure 7. Simulation results on (**a**) voice, (**b**) video, (**c**) CBR, and (**d**) system throughput when varying the number of SeNBs.

7. Conclusions

In this paper, we propose QDD and QED schemes to make dispatching decisions and to schedule radio resources, respectively. In QDD, the MeNB solves the designed linear programming formulation to obtain the dispatching decision. After dispatching traffics, the MeNB and SeNBs will then adopt the QED method to arrange the transmission of downlink data. The simulation results indicate that, compared to the other four combinations (mentioned in the simulation section), the proposed scheme can increase system throughput from 9% to 32% and decrease the packet drop ratio from 3% to 25%. In addition, the simulation results also verify that the proposed scheme can indeed preserve the QoS requirements of GBR traffics and increase overall system throughput. Since UEs are mobile in the real network, the proposed scheme does not consider how to effectively relay downlink data between the MeNB and SeNBs according to UEs' movements. Thus, in the future, we plan to design a handover prediction scheme to facilitate data dispatching.

Author Contributions: Conceptualization, M.-S.P.; methodology, M.-S.P. and H.-W.M.; software, H.-W.M.; investigation, H.S.; writing—review and editing, M.-S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work is sponsored by MOST 111-2628-E-027-002 and NTUT-BJUT Joint Research Program NTUT-BJUT-111-05.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sodhro, A.H.; Zahid, N. AI-enabled framework for fog computing driven e-healthcare applications. *Sensors* **2021**, *21*, 8039. [CrossRef] [PubMed]
2. Sukhmani, S.; Sadeghi, M.; Erol-Kantarci, M.; El Saddik, A. Edge caching and computing in 5G for mobile AR/VR and tactile internet. *IEEE Multimed.* **2018**, *26*, 21–30. [CrossRef]
3. LTE; E-UTRA and E-UTRAN; Overall Description (36.300). *3GPP Technical Specification (TS)*; 3GPP Specifications: Valbonne, France, 2017.
4. Antonioli, R.P.; Guerreiro, I.M.; Sousa, D.A.; Rodrigues, E.B.; Maciel, T.F.; Cavalcanti, F.R.P. User-assisted bearer split control for dual connectivity in multi-RAT 5G networks. *Wirel. Netw.* **2020**, *26*, 3675–3685. [CrossRef]
5. Ba, X. QoS-forecasting-based intelligent flow-control scheme for multi-connectivity in 5G heterogeneous networks. *IEEE Access* **2021**, *9*, 104304–104315. [CrossRef]
6. Cho, C.-W.; Pan, M.-S. Downlink radio resource scheduling for OFDMA systems with hybrid beamforming. *Wirel. Netw.* **2022**, *28*, 273–286. [CrossRef]
7. Ferdosian, N.; Othman, M.; Ali, B.M.; Lun, K.Y. Fair-QoS broker algorithm for overload-state downlink resource scheduling in LTE networks. *IEEE Syst. J.* **2017**, *12*, 3238–3249. [CrossRef]
8. Gemici, Ö.F.; Hokelek, I.; Cirpan, H.A. Trade-off analysis of QoS-aware configurable LTE downlink schedulers. In Proceedings of the IEEE International Conference on Telecommunications (ICT) 2013, Casablanca, Morocco, 6–8 May 2013.
9. Hong, E.K.; Baek, J.Y.; Jang, Y.O.; Na, J.H.; Kim, K.S. QoS-guaranteed scheduling for small cell networks. *ICT Express* **2018**, *4*, 175–180. [CrossRef]
10. Cui, H.; You, F. User-centric resource scheduling for dual-connectivity communications. *IEEE Commun. Lett.* **2021**, *25*, 3659–3663. [CrossRef]
11. Wu, Y.; Qian, L.P. Energy-efficient noma-enabled traffic offloading via dual-connectivity in small-cell networks. *IEEE Commun. Lett.* **2017**, 21, 1605–1608. [CrossRef]
12. Singh, S.; Geraseminko, M.; Yeh, S.P.; Himayat, N.; Talwar, S. Proportional fair traffic splitting and aggregation in heterogeneous wireless networks. *IEEE Commun. Lett.* **2016**, *20*, 1010–1013. [CrossRef]
13. Taksande, P.K.; Roy, A.; Karandikar, A. Optimal traffic splitting policy in LTE-based heterogeneous network. In Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC) 2018, Barcelona, Spain, 15–18 April 2018.
14. Wang, H.; Rosa, C.; Pedersen, K.I. Dual connectivity for LTE-advanced heterogeneous networks. *Wirel. Netw.* **2016**, *22*, 1315–1328. [CrossRef]
15. Prasad, A.; Maeder, A. Backhaul-aware energy efficient heterogeneous networks with dual connectivity. *Telecommun. Syst.* **2015**, *59*, 25–41. [CrossRef]
16. Pan, M.S.; Lin, T.M.; Chiu, C.Y.; Wang, C.Y. Downlink traffic scheduling for LTE-A small cell networks with dual connectivity enhancement. *IEEE Commun. Lett.* **2016**, *20*, 796–799. [CrossRef]
17. He, M.; Hua, C.; Xu, W.; Gu, P.; Shen, X.S. Delay optimal concurrent transmissions with raptor codes in dual connectivity networks. *IEEE Trans. Netw. Sci. Eng.* **2021**, *8*, 1478–1491. [CrossRef]
18. Sun, J.; Zhang, S.; Xu, S.; Cao, S. High throughput and low complexity traffic splitting mechanism for 5G non-stand alone dual connectivity transmission. *IEEE Access* **2021**, *9*, 65162–65172. [CrossRef]
19. Wang, C.; Huang, Y.-C. Delay-scheduler coupled throughput-fairness resource allocation algorithm in the long-term evolution wireless networks. *IET Commun.* **2014**, *8*, 3105–3112. [CrossRef]
20. Sandrasegaran, K.; Ramli, H.A.M.; Basukala, R. Delay-prioritized scheduling (DPS) for real time traffic in 3GPP LTE system. In Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC) 2010, Sydney, NSW, Australia, 18–21 April 2010.
21. Madi, N.K.; Hanapi, Z.M.; Othman, M.; Subramaniam, S.K. Delay-based and QoS-aware packet scheduling for RT and NRT multimedia services in LTE downlink systems. *EURASIP J. Wirel. Commun. Netw.* **2018**, *2018*, 1–21. [CrossRef]
22. Chaudhuri, S.; Baig, I.; Das, D. A novel QoS aware medium access control scheduler for LTE-advanced network. *Comput. Netw.* **2018**, *135*, 1–14. [CrossRef]
23. Nasralla, M.M. A hybrid downlink scheduling approach for multi-traffic classes in LTE wireless systems. *IEEE Access* **2020**, *8*, 82173–82186. [CrossRef]
24. Van den Eynde, J.; Blondia, C. A minimal delay violation downlink LTE scheduler. In Proceedings of the IEEE International Conference on Local Computer Networks (LCN) 2021, Edmonton, AB, Canada, 4–7 October 2021.
25. Kesavan, D.; Periyathambi, E.; Chokkalingam, A. A proportional fair scheduling strategy using multiobjective gradient-based african buffalo optimization algorithm for effective resource allocation and interference minimization. *Int. J. Commun. Syst.* **2022**, *35*, e5003. [CrossRef]
26. Wang, Y.-C.; Hsieh, S.-Y. Service-differentiated downlink flow scheduling to support QoS in long term evolution. *Comput. Netw.* **2016**, *94*, 344–359. [CrossRef]

27. QoS Class Identifier. Available online: https://en.wikipedia.org/wiki/QoS_Class_Identifier (accessed on 1 September 2022).
28. Kushner, H.J.; Whiting, P.A. Convergence of proportional-fair sharing algorithms under general conditions. *IEEE Trans. Wirel. Commun.* **2004**, *3*, 1250–1259. [CrossRef]
29. Lpsolver: Mixed Integer Linear Programming (MILP) Solver. Available online: http://lpsolve.sourceforge.net/5.5/ (accessed on 1 September 2022).