IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 13, 2019, accepted June 1, 2019, date of publication June 5, 2019, date of current version June 19, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2921075

Resource Allocation for Performance Enhancement in Mobile Ad Hoc Networks

JIA LIU[1,2], (Member, IEEE), YANG XU[3], (Member, IEEE), AND ZHAO LI[4], (Member, IEEE)

[1]State Key Laboratory of Integrated Services Network, Xidian University, Xi'an 710071, China
[2]Center for Cybersecurity Research and Development, National Institute of Informatics, Tokyo 101-8430, Japan
[3]School of Economics and Management, Xidian University, Xi'an 710071, China
[4]School of Cyber Engineering, Xidian University, Xi'an 710071, China

Corresponding author: Yang Xu (yxu@xidian.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61802292, in part by the National Key Research and Development Plan of China under Grant 2017YFB1400704, in part by the Project of Cyber Security Establishment with Inter University Cooperation, and in part by the Secom Science and Technology Foundation.

ABSTRACT The low-cost and easy-deployable mobile ad hoc networks (MANETs) have great potential to serve as the underlay network architecture for implementing lots of critical applications. In order to support different demands in such applications, understanding the fundamental performance of the MANETs and exploring their performance enhancement are of great importance. To this end, this paper, for the first time, investigates the performance enhancement of the MANETs by appropriately allocating the network resource. Specifically, we consider both the transmission resource and storage resource of a network node and use a two-tuple to depict a general resource allocation configuration. First, for a MANET with any given resource allocation configuration, we establish an analytical framework to model the network dynamics as queuing processes. With the help of the network modeling, we then derive the expressions of fundamental performance metrics, including per node throughput, expected an end-to-end delay, and throughput capacity. Based on these results, we further develop optimal resource allocation algorithms to determine the optimal setting of resource allocation, for enhancing network performance. Finally, we provide extensive simulation and numerical results to verify the efficiency of our theoretical modeling and illustrate the effects of resource allocation on network performance.

INDEX TERMS Mobile ad hoc networks, resource allocation, performance modeling, optimization.

I. INTRODUCTION

The rapid evolution of communication and network technologies has spawned the emergence of mobile ad hoc networks (MANETs). MANETs, composed of mobile nodes communicating with each other via wireless links without infrastructure and centralized administration, represent a kind of self-autonomous network architecture [1]. Due to the advantages of low-cost and easy-deployable, MANETs have great potential to serve as the underlay network architecture for implementing a wide range of promising applications, such as environmental monitoring, emergency rescue, smart homes, e-health, and so on [2], [3]. In order to support different demands in such applications, understanding the fundamental performance of MANETs (i.e., throughput, delay, capacity, etc.) is of great importance [4], [5].

Since the pioneering work of Grossglauser and Tse [6], extensive research efforts have been devoted to the modeling and performance evaluation of MANETs under various network scenarios. Specifically, Grossglauser and Tse studied the scaling law of per node throughput (i.e., the asymptotic behavior of per node throughput as the number of network nodes tends to infinity), and verified that a $\Theta(1)$ per node throughput can be achieved in MANET with the two-hop relay (2HR) routing scheme and independent and identically distributed (i.i.d) mobility model. Later, Neely and Modiano [7] derived the exact throughput capacity of MANETs and demonstrated that the lower bound of achievable delay-to-throughput ratio is $O(n)$ (n is the number of network nodes). Following this line, El Gamal *et al.* [8] and

The associate editor coordinating the review of this manuscript and approving it for publication was Mahdi Zareei.

Sharma *et al.* [9] then explored the throughput-delay tradeoff in MANETs under symmetric random walk mobility model and unified mobility model, respectively. Wang *et al.* further studied the throughput of MANETs with multicast traffic in [10], [11], and conducted performance comparison between the unicast and multicast MANETs in [12]. More recently, taking into account the fact that the resource (especially the storage resource) of a node in a MANET is limited, the modeling and performance evaluation for buffer-limited MANETs was conducted in [13]–[15].

While the above works represent significant progress in the performance study of MANETs, they only focused on the performance analysis under given network scenarios, but failed to address the problem that how to enhance the network performance under these scenarios. It is worth noting that the resource of a node in a general MANET is limited, thus the way of utilizing the resource will greatly affect the network performance. Allocating appropriately the resource in a MANET is very promising for its performance enhancement, which is beneficial to implementing a wide range of practical applications more efficiently. However, the research with this consideration is largely uninvestigated.

As a first effort to fill this gap, this paper focuses on a typical 2HR MANET which is widely used in previous studies, and study how to achieve performance enhancement by resource allocation. Observe that the role of a node in a MANET is not only a source and destination, it also serves as a relay to temporarily store and forward packets for other nodes, which will consume the storage resource and transmission resource of this node. Motivated by this, we consider that the storage resource of a network node can be flexibly allocated to store its own packets or other nodes' packets, and also the transmission resource can be flexibly allocated to transmit its own packets or help other nodes forward their packets. We use a two-tuple (η, κ) to depict a general resource allocation configuration, and explore how to optimize the network performance over the setting of (η, κ).

The main contributions of this paper are summarized as follows:

- We, for the first time, explore the performance enhancement of a MANET by allocating its storage and transmission resource. For a general resource allocation configuration (η, κ), we establish an analytical framework to model the network dynamics as queuing processes. With the help of the network modeling, we derive the exact expressions of fundamental performance metrics, including per node throughput, expected end-to-end (E2E) delay and throughput capacity.
- Based on the results under the general resource allocation configuration, we analyze the properties of network performance expressions, and further develop efficient algorithms to determine the optimal setting of (η, κ), for maximizing per node throughput, throughput capacity, and minimizing expected E2E delay, respectively.
- Finally, we conduct simulations to demonstrate the validity of our theoretical modeling and performance evaluation. Also, we provide extensive numerical results to illustrate how resource allocation affects network performance, which can serve as important guidelines for the configuration and operation of practical MANETs.

The remainder of this paper is organized as follows. We present some related works in Section II. Section III introduces the preliminaries involved in this study. We establish the network modeling in Section IV and derive the performance results in Section V. Section VI proposes the resource allocation algorithms, and Section VII shows the simulation and numerical results. Finally, we conclude this paper in Section VIII.

II. RELATED WORKS

A substantial amount of works have been devoted to the study of resource allocation in various types of wireless networks. For delay tolerant networks, Lee *et al.* [16] explored the optimal resource allocation scheme by jointly optimizing the link scheduling, routing and replication. In [17], Sun and Wang proposed a resource allocation algorithm to save power consumption in an OFDM-based heterogeneous network while satisfying all users' rate requirements. Wu *et al.* considered a wireless powered communication network, and investigated the energy efficiency maximization problem via joint time allocation and power control in [18]. Wang *et al.* presented in [19] a decentralized algorithm based on alternating direction method of multipliers, for computation offloading, resource allocation and Internet content caching optimization in heterogeneous wireless cellular networks with mobile edge computing. In [20], Liang *et al.* studied optimal spectrum sharing and power allocation algorithm to maximize the ergodic capacity for D2D-enabled vehicular networks. For the studies of resource allocation in cognitive radio networks, please kindly refer to the surveys in [21], [22].

For wireless sensor networks which have many similar features with ad hoc networks, there also exist several works studying the resource allocation problems. Specifically, Giannecchini *et al.* [23] proposed a fast online resource allocation algorithm to dynamically reconfigure a sensor network for accomplishing end-to-end sensing tasks more efficiently. Yuan and Yu [24] developed a cross-layer optimization framework to implement joint source coding, routing and power allocation in wireless sensor networks. Wang *et al.* [25] considered a small-scale wireless sensor network over fading TDMA channels, and developed optimal energy-efficient resource allocation schemes to minimize average users' power. More recently, resource allocation schemes for minimizing energy consumption and maximizing transmission rate in energy harvesting wireless sensor networks were explored in [26] and [27], respectively.

It is worth mentioning that the resource in the above studies usually refers to the radio spectrum, link scheduling and time, which is similar to the transmission resource involved in our paper. However, our work is distinguishable from the

existing ones in the sense that a different wireless environment, i.e., MANET scenario, is investigated. Moreover, the distinct feature of storage limitation in a MANET is also taken into account for resource allocation. All these differences clarify the novelty of this work.

III. PRELIMINARIES

A. SYSTEM MODELS

1) NETWORK MODEL

We consider a time-slotted cell-partitioned network model which is widely used in previous studies [7]–[10], [12]–[15], where a torus network area is evenly partitioned into c non-overlapping cells and n ($n \geq 3$) mobile nodes randomly move in the network area following a "uniform type" mobility model [6]. Let $X_i(t)$ denote the location of i-th node at time slot t, then with the "uniform type" mobility model, the stochastic process $\{X_i(\cdot)\}$ is stationary and ergodic with stationary distribution uniform on the network area, and the trajectories of different nodes are independent and identically distributed. It is worth noting that the "uniform type" mobility model is very general since it covers many typical mobility models as special cases, such as the i.i.d. mobility model [7], the random walk model [8], the random way-point model [28], and so on.

In a time slot, one cell supports only one transmission between a transmitter-receiver pair within it, the amount of data to be transmitted is fixed and normalized to one packet. Concurrent transmissions in different cells will not interfere with each other. As indicated in [7], wireless interference can be mitigated by requiring users in neighboring cells to transmit over orthogonal frequency bands. It is well known that for rectilinear cell partitioning, four separate frequency bands are sufficient for guaranteeing no neighboring cells use the same frequency, and this number can be reduced to three if cells are arranged in a hexagonal pattern. Moreover, additional frequency bands can be adopted to increase the frequency reuse distance.

We adopt the unicast traffic pattern with m traffic flows existing in the network. In this study, we consider the case of $m = n$, i.e., the permutation traffic model which is widely used in other works [6]–[8]. With the permutation traffic model, each node is the source of one traffic flow and meanwhile the destination of another traffic flow. More formally, let $\mathcal{D}(i)$ denote the destination node of the traffic flow originated from node i, then the source-destination pairs are matched in a way that the sequence $\{\mathcal{D}(1), \mathcal{D}(2), \cdots, \mathcal{D}(n)\}$ is just a derangement of the set of nodes $\{1, 2, \cdots, n\}$. Two typical examples are $\mathcal{D}(1) = 2, \mathcal{D}(2) = 3, \cdots, \mathcal{D}(n) = 1$, and $\mathcal{D}(1) = 2, \mathcal{D}(2) = 1, \mathcal{D}(3) = 4, \mathcal{D}(4) = 3, \cdots, \mathcal{D}(n-1) = n, \mathcal{D}(n) = n - 1$ (here n should be even). We assume that the packet generating process at each node is a Bernoulli process with mean rate λ_g, i.e., a new packet is generated by a node at the beginning of a time slot with probability λ_g. The storage space of a network node is limited and of size b packets, for storing its self-generated packets and packets from other nodes.

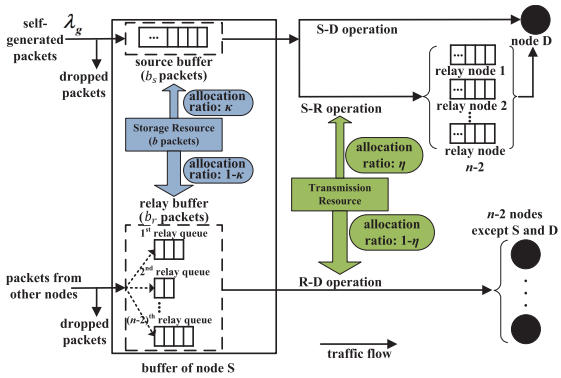


FIGURE 1. Illustration of resource allocation.

2) ROUTING SCHEME

To support the efficient operation of a MANET, we consider the popular two-hop relay routing (2HR) scheme [6], [7], [13]–[15]. Without loss of generality, we focus on a cell containing at least two nodes in a time slot, and the 2HR scheme is executed as follow.

1) If there exist source-destination pairs within the cell, equally likely choose such a pair to execute the source-to-destination (S-D) operation.
   - S-D operation: If the source contains packets intended for that destination, it transmits the earliest generated packet; else, it remains idle.
2) If there is no source-destination pair in the cell, equally likely designate a node within the cell as the transmitter, independently choose another node as the receiver uniformly over the remaining nodes. This transmitter-receiver pair executes the Source-to-Relay (S-R) or Relay-to-Destination (R-D) operation.
   - S-R operation: If the transmitter has self-generated packets, it relays the earliest generated packet to the designated receiver; else it remains idle.
   - R-D operation: If the transmitter has packets destined for the designated receiver, it forwards the oldest packet to the receiver; else it remains idle.

It is worth noting that with the help of node mobility, the 2HR relay scheme can guarantee the end-to-end packet delivery from a source to its destination.

B. RESOURCE ALLOCATION

1) STORAGE RESOURCE

As illustrated in Fig. 1, we consider that the storage resource of a node is allocated to its source buffer and relay buffer, separately. The source buffer is for storing the packets of its own flow (self-generated packets) and works as a FIFO (first-in-first-out) [29] source queue, while the relay buffer is for storing packets from all other $n - 2$ flows and works as $n - 2$ FIFO virtual relay queues (one queue per flow). With the FIFO rule, the early-arrival packet will be served early. When the buffer is fully occupied, the new arrival packet at the buffer will be dropped directly.

We use a parameter κ ($\frac{1}{b} \leq \kappa < 1$) to denote the ratio of storage resource allocated to the source buffer, and let b_s and b_r denote the source buffer size and relay buffer size, respectively, then we have

$$b_s = \lfloor \kappa \cdot b \rfloor, \tag{1}$$
$$b_r = \lceil (1-\kappa) \cdot b \rceil. \tag{2}$$

It can be seen that κ reflects the degree of selfishness of a node for the storage resource, i.e., κ is larger, then the node is more selfish.

2) TRANSMISSION RESOURCE

According to the routing scheme, in a time slot a node can transmit a packet only when it is designated as a transmitter, indicating the transmission resource of a node is also limited. When a node gains a precious transmission opportunity and its destination is not within the same cell, this transmission opportunity can be used to execute an S-R or R-D operation. As illustrated in Fig. 1, we use a parameter η ($0 < \eta < 1$) to denote the ratio of transmission resource allocated to the S-R operation. As in [7], let p and q denote the probabilities of finding at least two nodes in a cell, and finding at least one source-destination pair in a cell, respectively, then we have

$$p = 1 - \left(1 - \frac{1}{c}\right)^n - \frac{n}{c}\left(1 - \frac{1}{c}\right)^{n-1} \tag{3}$$
$$q = 1 - \left(1 - \frac{1}{c^2}\right)^{n/2}. \tag{4}$$

Let p_{sd}, p_{sr} and p_{rd} represents the probabilities that a node is scheduled to conduct an S-D, S-R and R-D operation, respectively, then we have the following system of linear equations:

$$\begin{cases} n \cdot (p_{sd} + p_{sr} + p_{rd}) = c \cdot p, \\ n \cdot p_{sd} = c \cdot q \\ \dfrac{p_{sr}}{p_{rd}} = \dfrac{\eta}{1-\eta} \triangleq \chi. \end{cases} \tag{5}$$

Solving this system of linear equations we have

$$p_{sd} = \frac{q}{d}, \tag{6}$$
$$p_{sr} = \frac{\eta(p-q)}{d}, \tag{7}$$
$$p_{rd} = \frac{(1-\eta)(p-q)}{d}, \tag{8}$$

where $d = \frac{n}{c}$ represents the density of the MANET. It can be seen that η reflects the degree of selfishness of a node for the transmission resource, i.e., η is larger, then the node is more selfish.

Therefore, the resource allocation configuration can be represented as a two-tuple (η, κ). In the following sections, we first conduct the network modeling and analyze the performance under a general (η, κ), and then find the optimal configuration for performance enhancement. The main notations in this paper is presented in Table 1.

TABLE 1. Main notations.

Symbol	Definition
n	Number of network nodes
c	Number of network cells
d	Density of the network
λ_g	Packet generating rate
b	Storage space of a node
κ	Storage resource allocation ratio
η	Transmission resource allocation ratio
χ	$\chi \triangleq \frac{\eta}{1-\eta}$
p	Probability of finding at least two nodes in a cell
q	Probability of finding at least one source-destination pair in a cell
p_d, p_{sr}, p_{rd}	Probabilities that a node is scheduled to conduct a S-D, S-R and R-D operation, respectively
μ_s	Service rate of the source queue
$\mathbf{\Pi}$	Stationary occupancy state distribution of the source buffer
Π_e	Probability that the source buffer is empty
$\mathbf{\Phi}$	Stationary occupancy state distribution of the relay buffer
Φ_f	probability that the relay buffer is full
T	Per node throughput
$\overline{D}_{e2e}$	Expected end-to-end delay
T_c	Throughput capacity

IV. MODELING OF BUFFER OCCUPANCY PROCESSES

In this section, for a general resource allocation configuration (η, κ), we establish an analytical framework based on the queuing theory and birth-death chain model, to fully characterize the source/relay buffer occupancy process. Due to the symmetry of network nodes and traffic flows, we only focus on a general node S in the following analysis.

A. MODELING OF SOURCE BUFFER OCCUPANCY PROCESS

Regarding the source buffer of node S, in every time slot a new packet is generated with probability λ_g, and a service opportunity arises with probability μ_s which can be determined as

$$\mu_s = p_{sd} + p_{sr} = \frac{\eta p + (1-\eta)q}{d}, \tag{9}$$

thus the occupancy process of source buffer can be modeled by a Bernoulli/Bernoulli/1/b_s queue as illustrated in Fig. 2.

Let Π_i denote the probability that there are i packets occupying the source buffer in the stationary state, then the stationary occupancy state distribution (OSD) of the source

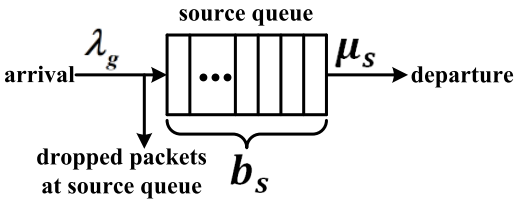


FIGURE 2. Bernoulli/Bernoulli/1/b_s queuing model for source buffer.

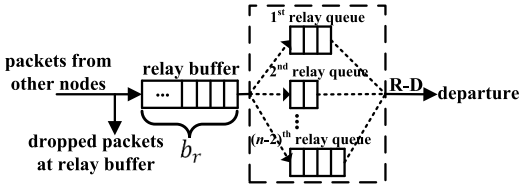


FIGURE 3. Illustration of packet occupancy process of the relay buffer.

buffer $\Pi = \left[\Pi_0, \Pi_1, \cdots, \Pi_{\lfloor \kappa b \rfloor}\right]$ is given by [30]

$$\Pi_i = \begin{cases} \dfrac{\mu_s - \lambda_g}{\mu_s - \lambda_g \cdot \tau^{\lfloor \kappa b \rfloor}}, & i = 0 \\ \dfrac{\mu_s - \lambda_g}{\mu_s - \lambda_g \cdot \tau^{\lfloor \kappa b \rfloor}} \dfrac{1}{1-\mu_s} \tau^i, & 1 \le i \le \lfloor \kappa b \rfloor \end{cases} \tag{10}$$

where $\tau = \dfrac{\lambda_g(1-\mu_s)}{\mu_s(1-\lambda_g)}$.

Let Π_e denote the probability that the source buffer is empty, then we have

$$\Pi_e = \Pi_0 = \frac{\mu_s - \lambda_g}{\mu_s - \lambda_g \cdot \tau^{\lfloor \kappa b \rfloor}}. \tag{11}$$

Let λ_d denote the packet departure rate of the source buffer, λ_d can be determined as

$$\lambda_d = \mu_s \cdot (1 - \Pi_e). \tag{12}$$

B. MODELING OF RELAY BUFFER OCCUPANCY PROCESS

Now we analyze the occupancy process of the relay buffer of node S, as illustrated in Fig. 3. Let X_t denote the number of packets in the relay buffer at time slot t, then the occupancy process of the relay buffer can be regarded as a stochastic process $\{X_t, t = 0, 1, 2, \cdots\}$ on the state space $\{0, 1, \cdots, \lceil (1-\kappa)b \rceil\}$. Notice that S cannot conduct an R-D operation and receive a packet from another node in the same time slot. Therefore, suppose that the relay buffer is at state j in the current time slot, only one of the following transition scenarios may happen in the next time slot:

- j to $j+1$ ($0 \le j < \lceil (1-\kappa)b \rceil$): the relay buffer is not full, and a packet arrives at the relay buffer.
- j to $j-1$ ($0 < j \le \lceil (1-\kappa)b \rceil$): the relay buffer is not empty, and a packet departs from the relay buffer.
- j to j ($0 \le j \le \lceil (1-\kappa)b \rceil$): no packet arrives at and departs from the relay buffer.

Let $p_{j,k}$ denote the one-step transition probability from state j to state k, then the occupancy process $\{X_t, t = 0, 1, 2, \cdots\}$ can be modeled by a birth-death chain as illustrated in Fig. 4. Let Φ_j denote the probability that there

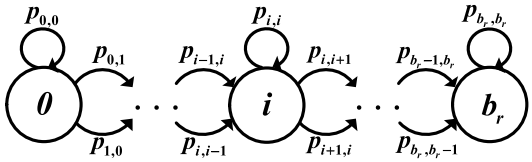


FIGURE 4. State machine of the birth-death chain.

are j packets occupying the relay buffer in the stationary state, the stationary OSD of the relay buffer $\Phi = \left[\Phi_0, \Phi_1, \cdots, \Phi_{\lceil (1-\kappa)b \rceil}\right]$ can be determined as

$$\Phi \cdot \mathbf{P} = \Phi, \tag{13}$$

$$\Phi \cdot \mathbf{1} = 1, \tag{14}$$

where $\mathbf{P}$ is the one-step transition matrix of the birth-death chain and $\mathbf{1}$ is a column vector of size $(\lceil (1-\kappa)b \rceil + 1) \times 1$ with all elements being 1.

From Fig. 4 and expressions (13) and (14), it can be seen the key to deriving Φ is the one-step transition probabilities $p_{j,j+1}$ and $p_{j,j-1}$. We provide the following lemma to determine $p_{j,j+1}$ and $p_{j,j-1}$, the proof can be found in our previous work [14] and thus is omitted here.

Lemma 1: For the birth-death chain in Fig. 4, its one-step transition probabilities $p_{j,j+1}$ and $p_{j,j-1}$ are determined as

$$p_{j,j+1} = p_{sr} \cdot (1 - \Pi_e), \quad 0 \le j < \lceil (1-\kappa)b \rceil, \tag{15}$$

$$p_{j,j-1} = p_{rd} \cdot \frac{j}{n-3+j}, \quad 0 < j \le \lceil (1-\kappa)b \rceil. \tag{16}$$

Substituting (15) and (16) into (13) and (14), the stationary OSD of the relay buffer is given by

$$\Phi_j = \frac{\mathrm{C}_j((1-\Pi_e)\chi)^j}{\sum_{k=0}^{\lceil (1-\kappa)b \rceil} \mathrm{C}_k((1-\Pi_e)\chi)^k}, \quad 0 \le j \le \lceil (1-\kappa)b \rceil \tag{17}$$

where $\mathrm{C}_j = \binom{n-3+j}{j}$. Let Φ_f denote the probability that the relay buffer is full, then we have

$$\Phi_f = \Phi_{\lceil (1-\kappa)b \rceil} = \frac{\mathrm{C}_{\lceil (1-\kappa)b \rceil}((1-\Pi_e)\chi)^{\lceil (1-\kappa)b \rceil}}{\sum_{k=0}^{\lceil (1-\kappa)b \rceil} \mathrm{C}_k((1-\Pi_e)\chi)^k}. \tag{18}$$

V. PERFORMANCE EVALUATION

With the help of network modeling, in this section we derive the exact expressions of fundamental performance metrics, including per node throughput, expected E2E delay and throughput capacity.

A. PER NODE THROUGHPUT

We use T to denote the per node throughput, which is defined as the time-average number of packets that can be delivered from a source to its destination.

Notice that the packets destined for a destination node can be delivered to this destination through either one-hop

transmission (S-D) or two-hop transmission (S-R and R-D). Thus, we can decompose the per node throughput T into the packet delivery rates through these two ways of transmission. Let T_{sd} and T_{rd} denote the packet delivery rates at the destination node through its source node and relay nodes, respectively, then we have

$$T_{sd} = \lambda_d \cdot \frac{p_{sd}}{\mu_s}, \tag{19}$$

$$T_{rd} = \lambda_d \cdot \frac{p_{sr}}{\mu_s} \cdot \left(1 - \Phi_f\right). \tag{20}$$

Therefore, the per node throughput T is given by

$$\begin{aligned} T &= T_{sd} + T_{rd} \\ &= \frac{q(1-\Pi_e) + \eta(p-q)(1-\Pi_e)(1-\Phi_f)}{d}, \end{aligned} \tag{21}$$

where p, q, Π_e and Φ_f are given by the expressions (3), (4), (11) and (18), respectively.

B. EXPECTED E2E DELAY

We use $\overline{D}_{e2e}$ to denote the expected E2E delay, which is defined as the expected time it takes a packet to reach its destination after it is generated by its source. It should be mentioned that for the calculation of $\overline{D}_{e2e}$, we only focus on the packets that have been successfully delivered to their destinations.

Let $\widetilde{\Pi}_i$ $(0 \le i \le \lfloor \kappa b \rfloor - 1)$ denote the probability that there are i packets in the source buffer conditioned on that the source buffer is not full. From [30] we have

$$\widetilde{\Pi}_i = \frac{1-\tau}{1-\tau^{\lfloor \kappa b \rfloor}} \tau^i. \tag{22}$$

Then, the conditional expectation of source queue length $\widetilde{L}_s$ can be determined as

$$\widetilde{L}_s = \sum_{i=0}^{\lfloor \kappa b \rfloor - 1} i\widetilde{\Pi}_i = \frac{\tau - \lfloor \kappa b \rfloor \tau^{\lfloor \kappa b \rfloor} + (\lfloor \kappa b \rfloor - 1)\tau^{\lfloor \kappa b \rfloor + 1}}{(1-\tau)(1-\tau^{\lfloor \kappa b \rfloor})}. \tag{23}$$

Thus, the expected queuing delay $\overline{Q}_s$ and total delay $\overline{D}_s$ in the source queue are given by

$$\overline{Q}_s = \frac{\widetilde{L}_s}{\mu_s}, \tag{24}$$

$$\overline{D}_s = \overline{Q}_s + \frac{1}{\mu_s}. \tag{25}$$

Let $\widetilde{\Phi}_j$ $(0 \le j \le \lceil (1-\kappa)b \rceil - 1)$ denote the probability that there are j packets in the relay buffer conditioned on that the relay buffer is not full. We have $\widetilde{\Phi}_j = \frac{\Phi_j}{1-\Phi_f}$ and the conditional expectation of the number of packets in relay buffer $\widetilde{L}_r$ is calculated as

$$\widetilde{L}_r = \sum_{j=0}^{\lceil (1-\kappa)b \rceil - 1} j\widetilde{\Phi}_j = \frac{\sum_{k=0}^{\lceil (1-\kappa)b \rceil - 1} k\mathrm{C}_k((1-\Pi_e)\chi)^k}{\sum_{k=0}^{\lceil (1-\kappa)b \rceil - 1} \mathrm{C}_k((1-\Pi_e)\chi)^k}. \tag{26}$$

Then, the expected queuing delay $\overline{Q}_r$ and total delay $\overline{D}_r$ in the relay queue can be determined as

$$\overline{Q}_r = \frac{\widetilde{L}_r}{n-2} \cdot \left(\frac{p_{rd}}{n-2}\right)^{-1}, \tag{27}$$

$$\overline{D}_r = \overline{Q}_r + \left(\frac{p_{rd}}{n-2}\right)^{-1}. \tag{28}$$

Therefore, the expected E2E delay $\overline{D}_{e2e}$ is given by

$$\overline{D}_{e2e} = \overline{D}_s + \overline{D}_r \cdot \frac{T_{rd}}{T_{sd} + T_{rd}} \tag{29}$$

$$= d\left\{\frac{1+\widetilde{L}_s}{\eta p + (1-\eta)q} + \frac{\chi(n-2+\widetilde{L}_r)(1-\Phi_f)}{q + \eta(p-q)(1-\Phi_f)}\right\}. \tag{30}$$

In addition, we can approximate $\widetilde{L}_r$ as

$$\begin{aligned} \widetilde{L}_r &= (n-2)\frac{(1-\Pi_e)\chi - ((1-\Pi_e)\chi)^{\lceil (1-\kappa)b \rceil}}{1-(1-\Pi_e)\chi} \\ &\quad + \mathcal{O}\left\{((1-\Pi_e)\chi)^{\lceil (1-\kappa)b \rceil}\right\} \end{aligned} \tag{31}$$

$$\approx \widehat{L}_r = (n-2)\frac{(1-\Pi_e)\chi - ((1-\Pi_e)\chi)^{\lceil (1-\kappa)b \rceil}}{1-(1-\Pi_e)\chi}. \tag{32}$$

Substituting (32) into (30) we can obtain an approximation of the expected E2E delay.

C. THROUGHPUT CAPACITY

We use T_c to denote the throughput capacity. For the homogeneous network scenario considered in this paper, the network level throughput capacity can be defined by the maximal achievable per node throughput over any packet generating rate λ_g, i.e., $T_c = \max_{\lambda_g \in (0,1]} T$.

For deriving the expression of T_c, we first need the following lemma.

Lemma 2: T monotonically increases with λ_g.

Proof: According to formula (11), the derivative of Π_e with respect to λ_g is

$$\begin{aligned} &\frac{\partial \Pi_e}{\partial \lambda_g} \\ &= \frac{-\mu_s + \mu_s \tau^{\lfloor \kappa b \rfloor} + \lfloor \kappa b \rfloor \frac{\mu_s - \lambda_g}{1-\lambda_g} \tau^{\lfloor \kappa b \rfloor}}{(\mu_s - \lambda_g \tau^{\lfloor \kappa b \rfloor})^2} \\ &= \frac{-(\lambda_g - \mu_s)^2}{(\mu_s - \lambda_g \tau^{\lfloor \kappa b \rfloor})^2 \cdot (1-\lambda_g)^2 \cdot \mu_s} \cdot \sum_{i=1}^{\lfloor \kappa b \rfloor} i\tau^{i-1} < 0. \end{aligned} \tag{33}$$

Thus, as λ_g increases, Π_e decreases.

Substituting (18) into (21), T can be expressed as

$$T = \frac{1-\Pi_e}{d}\left\{q + \eta(p-q)\frac{\sum_{j=0}^{\lceil (1-\kappa)b \rceil - 1} \mathrm{C}_j((1-\Pi_e)\chi)^j}{\sum_{j=0}^{\lceil (1-\kappa)b \rceil} \mathrm{C}_j((1-\Pi_e)\chi)^j}\right\}. \tag{34}$$

Let $\omega = \frac{1}{1-\Pi_e}$, $\delta = \frac{1}{\chi}$, $h(\omega) = \sum_{j=0}^{\lceil(1-\kappa)b\rceil-1} C_j(\omega\delta)^{\lceil(1-\kappa)b\rceil-j}$, and $g(\omega) = \omega \cdot \left(1 + \frac{C_{\lceil(1-\kappa)b\rceil}}{h(\omega)}\right)$, then T can be re-expressed as

$$T = \frac{1}{d}\left\{\frac{q}{\omega} + \frac{\eta(p-q)}{g(\omega)}\right\}. \tag{35}$$

Regarding the derivative of $g(\omega)$, we have

$$g'(\omega) = \frac{1}{h(\omega)^2} \cdot \underbrace{\left\{h(\omega)(h(\omega) + C_{\lceil(1-\kappa)b\rceil}) - \omega C_{\lceil(1-\kappa)b\rceil} h'(\omega)\right\}}_{(a)}, \tag{36}$$

and we derive expression (a) > 0, as shown at the bottom of the next page, where (37) holds because $C_{\lceil(1-\kappa)b\rceil-j} C_{\lceil(1-\kappa)b\rceil-k+j} > C_{\lceil(1-\kappa)b\rceil} C_{\lceil(1-\kappa)b\rceil-k}$ for $0 < j < k$.

In summary, as λ_g increases, Π_e decreases, then ω decreases, then $g(\omega)$ decreases, finally we have T increases. ■

With Lemma 2 we have

$$T_c = \max_{\lambda_g \in (0,1]} T = \lim_{\lambda_g \to 1} T. \tag{38}$$

From (11) and (18) we have

$$\lim_{\lambda_g \to 1} \tau = \lim_{\lambda_g \to 1} \frac{\lambda_g(1-\mu_s)}{\mu_s(1-\lambda_g)} \to \infty,$$

$$\lim_{\lambda_g \to 1} \Pi_e = \lim_{\lambda_g \to 1} \frac{\mu_s - \lambda_g}{\mu_s - \lambda_g \cdot \tau^{\lfloor \kappa b \rfloor}} = 0, \tag{39}$$

$$\lim_{\lambda_g \to 1} \Phi_f = \frac{C_{\lceil(1-\kappa)b\rceil}\chi^{\lceil(1-\kappa)b\rceil}}{\sum_{k=0}^{\lceil(1-\kappa)b\rceil} C_k \chi^k}. \tag{40}$$

Therefore, substituting (39) and (40) into (21), the throughput capacity T_c is given by

$$T_c = \frac{1}{d}\left\{q + \eta(p-q) \cdot \frac{\sum_{k=0}^{\lceil(1-\kappa)b\rceil-1} C_k \chi^k}{\sum_{k=0}^{\lceil(1-\kappa)b\rceil} C_k \chi^k}\right\}. \tag{41}$$

VI. OPTIMAL RESOURCE ALLOCATION

Based on the performance modeling results, in this section we explore the problems of optimal resource allocation. By allocating the network node resource (transmission resource and storage resource) appropriately, i.e., determining the optimal setting of (η, κ), we can optimize the desired performance metrics to satisfy various requirements in different practical applications.

A. THROUGHPUT MAXIMIZATION

We first consider the optimal resource allocation for throughput maximization, which can be mathematically formulated as the following optimization problem.

P1 (Throughput Maximization):

$$\max_{(\eta,\kappa)} \ T \tag{42a}$$

$$\text{s.t. } 0 < \eta < 1, \tag{42b}$$

$$\frac{1}{b} \le \kappa < 1. \tag{42c}$$

In order to solve problem (42), we exploit the technique of alternating optimization, which is a well-established methodology to optimize the objective function by first optimizing over some of the variables and then optimizing over the remain ones [31]. Therefore, we optimize the transmission resource allocation and storage source allocation in an alternating manner. Specifically, for a fixed κ, problem (42) can be re-formulated as

$$\max_{0 \le \eta \le 1} T = \frac{q(1-\Pi_e) + \eta(p-q)(1-\Pi_e)(1-\Phi_f)}{d}. \tag{43}$$

We can see from expressions (11) and (18) that it is hard to obtain the closed form expression of the optimal solution by calculating the derivative of T with respect to η. However, we can apply some efficient one-dimensional search methods. From (9) we know that μ_s increases as η increases, and from (11) we have

$$\begin{aligned}
&\frac{\partial \Pi_e}{\partial \mu_s} \\
&= \frac{\mu_s - \lambda_g\tau^{\lfloor \kappa b \rfloor} - \left(1 - \lambda_g \lfloor \kappa b \rfloor \tau^{\lfloor \kappa b \rfloor - 1} \frac{\partial \tau}{\partial \mu_s}\right)(\mu_s - \lambda_g)}{(\mu_s - \lambda_g \tau^{\lfloor \kappa b \rfloor})^2} \\
&= \frac{\lambda_g - \lambda_g \tau^{\lfloor \kappa b \rfloor} - \lfloor \kappa b \rfloor \frac{\lambda_g(\mu_s - \lambda_g)}{\mu_s(1-\mu_s)} \tau^{\lfloor \kappa b \rfloor}}{(\mu_s - \lambda_g \tau^{\lfloor \kappa b \rfloor})^2} \\
&= \frac{\lambda_g(\mu_s - \lambda_g)^2}{\mu_s^2(\mu_s - \lambda_g\tau^{\lfloor \kappa b \rfloor})^2(1-\lambda_g)} \sum_{i=0}^{\lfloor \kappa b \rfloor - 1} \left(1 + \frac{i}{1-\mu_s}\right)\tau^i \\
&> 0,
\end{aligned} \tag{44}$$

which indicates Π_e also increases as η increases. Similarly, we can verify that $\frac{\partial \Phi_f}{\partial \eta} > 0$.

According to the structure of objective function (43), we can conclude that T is either a unimodal function or monotonic function of η, inspiring us to adopt the golden-section search method to find the optimal solution η^*. Therefore, the optimal transmission resource allocation algorithm for maximizing the per node throughput can be summarized in Algorithm 1. Note that with Algorithm 1, the search interval will be reduced to its 0.618 times after each iteration. The initial search interval is 1 and the convergence tolerance is ε. Thus, we need $\lceil \log_{0.618} \varepsilon + 1 \rceil$ iterations to reduce the search interval to being less than ε (i.e., the algorithm stops). We also set the maximum number of iterations to be v_{max}. Therefore, the number of iterations can be estimated as $\min\{\lceil \log_{0.618} \varepsilon + 1 \rceil, v_{max}\}$.

Next, we consider that for a fixed η, how to maximize the throughput by allocating the storage resource. From (1)

Algorithm 1 Optimal Transmission Resource Allocation Algorithm for Maximizing Throughput

Input: Network parameters $(n, c, b, \kappa, \lambda_g)$;
Output: Optimal transmission resource allocation ratio η^*;

Initialization: set the convergence tolerance $\varepsilon > 0$, the maximum number of iterations v_{max}, the iteration index $v = 0$, $Z_l = 0$, $Z_u = 1$, $x_1(v) = 0.382$ and $x_2(v) = 0.618$;
repeat
  $y_1 = T|_{\eta=x_1(v)}$;
  $y_2 = T|_{\eta=x_2(v)}$;
  if $y_1 < y_2$ **then**
    $Z_l = x_1(v), Z_u = Z_u$;
    $x_1(v+1) = x_2(v)$;
    $x_2(v+1) = Z_l + 0.618(Z_u - Z_l)$;
  else
    $Z_l = Z_l, Z_u = x_2(v)$;
    $x_2(v+1) = x_1(v)$;
    $x_1(v+1) = Z_l + 0.382(Z_u - Z_l)$;
  $v = v + 1$;
until $|x_1(v) - x_2(v)| < \varepsilon$ *or* $v > v_{max}$;
$\eta^* = x_1(v)$;
return η^*;

and (2), problem (42) can be re-formulated as

$$\max_{b_s\in\{1,2,\cdots,b-1\}} T = \frac{q(1-\Pi_e)+\eta(p-q)(1-\Pi_e)(1-\Phi_f)}{d}, \tag{45}$$

which is an integer programming.

From (11) we have

$$\Pi_e|_{b_s=t+1} - \Pi_e|_{b_s=t} = \frac{\lambda_g\tau^t(\mu_s-\lambda_g)(\tau-1)}{(\mu_s-\lambda_g\tau^{t+1})(\mu_s-\lambda_g\tau^t)} < 0. \tag{46}$$

From (18) we have $\frac{\partial\Phi_f}{\partial\Pi_e} < 0$. Moreover, we denote $\Xi \triangleq (1-\Pi_e)\chi$ and substitute it into (18), there is

$$\begin{aligned}
&\Phi_f|_{b_r=l+1} - \Phi_f|_{b_r=l} \\
&= \frac{\mathrm{C}_{l+1}\Xi^{l+1}}{\sum_{i=0}^{l+1}\mathrm{C}_i\Xi^i} - \frac{\mathrm{C}_l\Xi^l}{\sum_{i=0}^{l}\mathrm{C}_i\Xi^i} \\
&= \frac{\underbrace{\mathrm{C}_{l+1}\Xi^{l+1}\sum_{i=0}^{l}\mathrm{C}_i\Xi^i - \mathrm{C}_l\Xi^l\sum_{i=0}^{l+1}\mathrm{C}_i\Xi^i}_{(b)}}{\sum_{i=0}^{l+1}\mathrm{C}_i\Xi^i\cdot\sum_{i=0}^{l}\mathrm{C}_i\Xi^i},
\end{aligned}$$

and

$$\begin{aligned}
(b) &= \mathrm{C}_{l+1}\Xi^{l+1}\sum_{i=0}^{l}\mathrm{C}_i\Xi^i - \mathrm{C}_l\Xi^l\sum_{i=1}^{l+1}\mathrm{C}_i\Xi^i - \mathrm{C}_l\Xi^l \\
&< \sum_{i=0}^{l}(\mathrm{C}_{l+1}\mathrm{C}_i - \mathrm{C}_l\mathrm{C}_{i+1})\,\Xi^{l+i+1} < 0.
\end{aligned}$$

Then we have

$$\Phi_f|_{b_s=t+1} - \Phi_f|_{b_s=t} > 0. \tag{47}$$

Therefore, we can conclude that T is also either a unimodal function or monotonic function of b_s. We propose the following search method to determine the optimal storage resource allocation, as summarized in Algorithm 2. The number of iterations in Algorithm 2 can be calculated as $\lceil \log_2\ b+1 \rceil$.

Based on Algorithm 1 and Algorithm 2, we further propose the alternative optimization algorithm for determining the optimal resource allocation configuration (η^*, b_s^*) and the corresponding optimal throughput T^*, as shown in Algorithm 3. It is worth mentioning that Algorithm 3 executes Algorithm 1 and Algorithm 2 alternatively, and it always converges because the value of the objective function is finite and non-decreasing at each iteration.

$$\begin{aligned}
(a) &= \sum_{j=0}^{\lceil(1-\kappa)b\rceil-1}\mathrm{C}_j(\omega\delta)^{\lceil(1-\kappa)b\rceil-j}\cdot\sum_{j=0}^{\lceil(1-\kappa)b\rceil}\mathrm{C}_j(\omega\delta)^{\lceil(1-\kappa)b\rceil-j} - \mathrm{C}_{\lceil(1-\kappa)b\rceil}\sum_{j=0}^{\lceil(1-\kappa)b\rceil-1}\mathrm{C}_j(\lceil(1-\kappa)b\rceil-j)(\omega\delta)^{\lceil(1-\kappa)b\rceil-j} \\
&= \sum_{k=1}^{\lceil(1-\kappa)b\rceil}\left\{\sum_{j=0}^{k-1}\left(\mathrm{C}_{\lceil(1-\kappa)b\rceil-j}(\omega\delta)^j\mathrm{C}_{\lceil(1-\kappa)b\rceil-k+j}(\omega\delta)^{k-j}\right) - k\mathrm{C}_{\lceil(1-\kappa)b\rceil}\mathrm{C}_{\lceil(1-\kappa)b\rceil-k}(\omega\delta)^k\right\} \\
&\quad + \sum_{k=\lceil(1-\kappa)b\rceil+1}^{2\lceil(1-\kappa)b\rceil}\sum_{j=k-\lceil(1-\kappa)b\rceil}^{\lceil(1-\kappa)b\rceil}\mathrm{C}_{B-j}(\omega\delta)^j\mathrm{C}_{B-k+j}(\omega\delta)^{k-j} \\
&> \sum_{k=1}^{\lceil(1-\kappa)b\rceil}\left\{\left\{\sum_{j=0}^{k-1}\left(\mathrm{C}_{\lceil(1-\kappa)b\rceil-j}\mathrm{C}_{\lceil(1-\kappa)b\rceil-k+j}\right) - k\mathrm{C}_{\lceil(1-\kappa)b\rceil}\mathrm{C}_{\lceil(1-\kappa)b\rceil-k}\right\}\omega^k\delta^k\right\} > 0,
\end{aligned} \tag{37}$$

Algorithm 2 Optimal Storage Resource Allocation Algorithm for Maximizing Throughput

Input: Network parameters $(n, c, b, \eta, \lambda_g)$;
Output: Optimal source buffer size b_s^*;
Initialization: set $Z_l = 1, Z_u = b - 1, t = Z_l + \lfloor \frac{Z_u - Z_l}{2} \rfloor$;
repeat
  $y_1 = T|_{b_s=t}$;
  $y_2 = T|_{b_s=t+1}$;
  if $y_1 < y_2$ **then**
    $Z_l = t, Z_u = Z_u$;
  else
    $Z_l = Z_l, Z_u = t + 1$;
  $t = Z_l + \lfloor \frac{Z_u - Z_l}{2} \rfloor$;
until $Z_u - Z_l \leq 2$;
$U(i) = T|_{b_s=t+i}, i = 0, 1, 2$;
$b_s^* = t + \arg\max_i \{U(i)\}$;
return b_s^*;

Algorithm 3 Optimal Resource Allocation Algorithm for Throughput Maximization

Input: Network parameters (n, c, b, λ_g);
Output: Optimal resource allocation configuration (η^*, b_s^*) and optimal throughput T^*;
Initialization: set the convergence tolerance $\xi > 0$, the maximum number of iterations w_{max}, the iteration index $w = 0$, and initialize the resource allocation configuration $(\eta(w), b_s(w))$, ;
repeat
  Given $b_s(w)$, update $\eta(w + 1)$ according to Algorithm 1 (**Optimal Transmission Resource Allocation Algorithm**);
  Given $\eta(w + 1)$, update $b_s(w + 1)$ according to Algorithm 2 (**Optimal Storage Resource Allocation Algorithm**);
  $w = w + 1$;
  Update $T(w)$ according to formula (21);
until $T(w) - T(w - 1) < \xi$ *or* $w > w_{max}$;
$\eta^* = \eta(w), b_s^* = b_s(w), T^* = T(w)$;
return (η^*, b_s^*) and T^*;

B. EXPECTED E2E DELAY MINIMIZATION

We then consider the optimal resource allocation for expected E2E delay minimization, which can be mathematically formulated as the following optimization problem.

P2 (Expected E2E Delay Minimization):

$$\min_{(\eta,\kappa)} \ \overline{D}_{e2e} \tag{48a}$$
$$\text{s.t. } 0 < \eta < 1, \tag{48b}$$
$$\frac{1}{b} \leq \kappa < 1. \tag{48c}$$

Similar to the procedures of solving problem P1, we also exploit the technique of alternating optimization to convert P2 into two one-dimensional searching problems, and then apply searching algorithms to obtain the optimal resource allocation for minimizing the expected E2E delay. The details are omitted here.

C. THROUGHPUT CAPACITY MAXIMIZATION

Finally, we consider the optimal resource allocation for throughput capacity maximization, which can be mathematically formulated as the following optimization problem.

P3 (Throughput Capacity Maximization):

$$\max_{(\eta,\kappa)} \ T_c \tag{49a}$$
$$\text{s.t. } 0 < \eta < 1, \tag{49b}$$
$$\frac{1}{b} \leq \kappa < 1. \tag{49c}$$

According to the expression of T_c in formula (41), we can easily use the similar derivation of (47) to verify that T_c always increases as b_r increases. Therefore, the optimal storage resource allocation for maximizing throughput capacity can be determined as $b_s^* = 1, b_r^* = b - 1$, i.e., $\kappa^* \in [\frac{1}{b}, \frac{2}{b})$. As a result, problem P3 reduces to the following one-dimensional optimization issue over η.

$$\max_{0<\eta<1} T_c = \frac{1}{d}\left\{ q + \eta(p-q) \cdot \frac{\sum_{k=0}^{b-2} \mathrm{C}_k \chi^k}{\sum_{k=0}^{b-1} \mathrm{C}_k \chi^k} \right\}. \tag{50}$$

In the following, we develop an efficient algorithm based on the novel fixed-point iteration technique [32] to determine the optimal transmission resource allocation ratio η^*. We let $\delta = \frac{1}{\chi} = \frac{1-\eta}{\eta}$ and construct two functions as $H(\delta) = \sum_{k=0}^{b-2} \mathrm{C}_k \delta^{b-1-k}$ and $G(\delta) = (1+\delta)\left(1 + \frac{\mathrm{C}_{b-1}}{H(\delta)}\right)$, then T_c can be expressed as

$$T_c = \frac{1}{d}\left(q + \frac{p-q}{G(\delta)}\right).$$

Thus, maximizing T_c is equivalent to minimizing $G(\delta)$. Regarding $G(\delta)$, we have the following observations: i) $G(\delta)$ is a differentiable function; ii) $\lim_{\delta\to 0} G(\delta) \to \infty$ and $\lim_{\delta\to\infty} G(\delta) \to \infty$. According to the Extreme Value Theorem [33], we can determine δ^* which enables $G(\delta)$ to achieve its minimum value as $G'(\delta^*) = 0$. Therefore, δ^* is the solution of the following equation.

$$H(\delta)[H(\delta) + \mathrm{C}_{b-1}] = \mathrm{C}_{b-1}(1+\delta)H'(\delta).$$

It is hard to obtain the closed-form solution of this equation. To address this issue, we construct a fixed-point function as follow:

$$F_P(\delta) = \frac{\mathrm{C}_{b-1}(1+\delta)H'(\delta)}{H(\delta)[H(\delta) + \mathrm{C}_{b-1}]} \cdot \delta. \tag{51}$$

$F_P(\delta)$ is well-designed since it has the good property of contraction mapping. Therefore, we can efficiently find δ^* by executing the fixed-point function iteratively. Moreover, it

Algorithm 4 Optimal Resource Allocation Algorithm for Maximizing Throughput Capacity

Input: Network parameters (n, c, b);
Output: Optimal resource allocation configuration (η^*, b_s^*) and optimal throughput capacity T_c^*;
Initialization: set the convergence tolerance $\varepsilon > 0$, the maximum number of iterations u_{max}, the iteration index $u = 0$, and initialize $\delta_u > 1$;
repeat
  $\delta_{u+1} = F_P(\delta_u)$;
  $u = u + 1$;
until $\delta_u - \delta_{u-1} < \varepsilon$ *or* $u > u_{max}$;
$\eta^* = \frac{1}{1+\delta_u}$, $b_s^* = 1$, $T_c^* = T_c|_{(\eta^*, b_s^*)}$;
return (η^*, b_s^*, T_c^*);

can be verified that $G'(\delta) < 0$ for $\delta \in (0, 1)$. Thus, we can set the initial iteration point $\delta_0 > 1$. Note that $\eta = \frac{1}{1+\delta}$, it implies that the optimal transmission resource allocation ratio always satisfies $\eta^* \leq 0.5$, i.e., a network node needs to allocate more transmission opportunities to forwarding packets of other nodes. The detailed algorithm for maximizing throughput capacity can be summarized in Algorithm 4.

VII. SIMULATION

In this section, we first conduct simulations to verify the efficiency of our analytical framework and performance evaluation. Then we provide numerical results to illustrate the effects of resource allocation on network performance.

A. SIMULATION SETTINGS

To verify the efficiency of our analytical framework and performance evaluation, we developed a dedicated C++ network simulator to simulate the packet generating, queuing and delivering processes in a concerned MANET [34]. All the basic network parameters $(n, c, b, \eta, \kappa, \lambda_g)$ can be set flexibly. Without loss of generality, we set two network configurations for the performance validation, Case 1: $(n = 30, c = 15, b = 10, \kappa = 0.35, \eta = 0.3)$, and Case 2: $(n = 50, c = 20, b = 15, \kappa = 0.5, \eta = 0.6)$. Each simulation task runs over a period of 2×10^8 time slots, and we only collect data from the last 80% of time slots to ensure the system is in the steady state. In each simulation task, we count the number of packets delivered to a tagged node and the delay of each packet (denoted by N and $D(i)$, respectively), and calculate the simulation results as

$$T|_{\text{simulation}} = \frac{N}{2 \cdot 10^8}, \tag{52}$$

$$\overline{D}_{e2e}|_{\text{simulation}} = \frac{\sum_{i=1}^{N} D(i)}{N}. \tag{53}$$

In the simulator, three typical specific mobility models which belong to the "uniform type" mobility model, i.e., i.i.d., random walk and random way-point, have been implemented. We only show the simulation results under the i.i.d. mobility model.

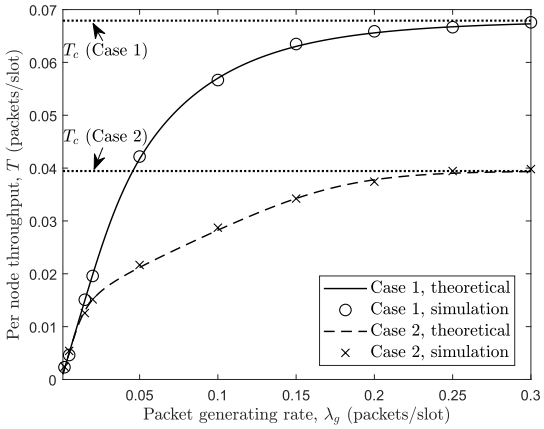


FIGURE 5. Validation for the per node throughput modeling. Case 1: $(n = 30, c = 15, b = 10, \kappa = 0.35, \eta = 0.3)$. Case 2: $(n = 50, c = 20, b = 15, \kappa = 0.5, \eta = 0.6)$.

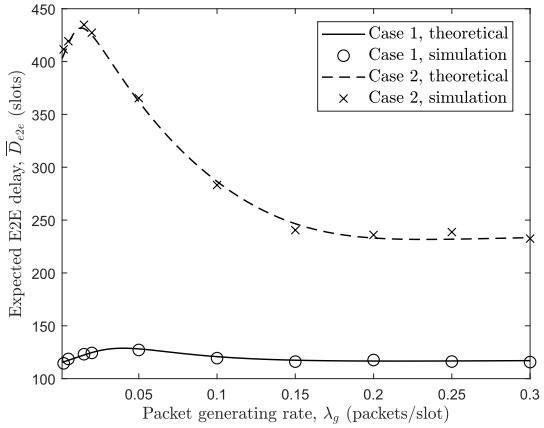


FIGURE 6. Validation for the expected E2E delay modeling. Case 1: $(n = 30, c = 15, b = 10, \kappa = 0.35, \eta = 0.3)$. Case 2: $(n = 50, c = 20, b = 15, \kappa = 0.5, \eta = 0.6)$.

B. VALIDATION

First, we summarize in Fig. 5 the theoretical and simulation results of per node throughput T versus packet generating rate λ_g under the two cases. We can see from Fig. 5 that both simulation results can match the corresponding theoretical curves very nicely, indicating that our analytical framework is highly efficient for modeling the throughput performance of MANETs under a general resource allocation configuration. In addition, as indicated in Lemma 2, we can observe that the per node throughput T increases monotonically as the packet generating rate λ_g increases, and finally tends to a constant which is the throughput capacity T_c.

We then plot Fig. 6 to show the theoretical and simulation results of expected E2E delay $\overline{D}_{e2e}$ versus λ_g under the two cases. Fig. 6 shows clearly that the simulation results match well with the corresponding theoretical curves for both the

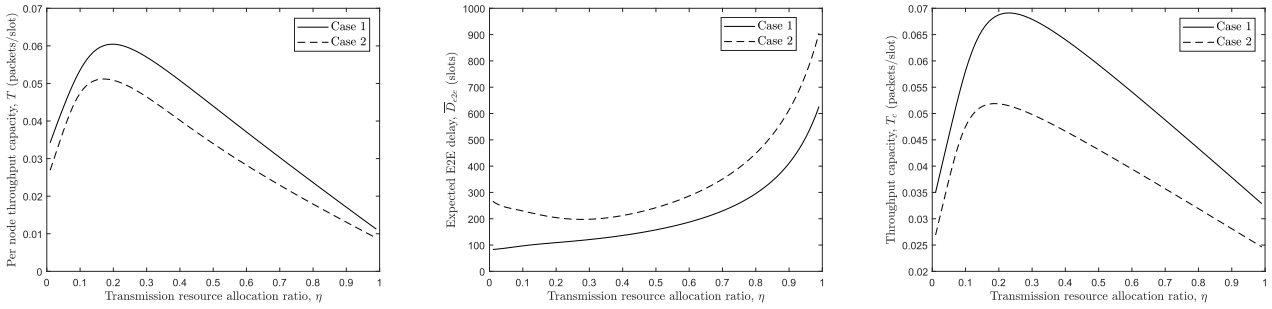


FIGURE 7. Effects of transmission resource allocation on network performance. Case 1: ($n = 30, c = 15, b = 10, \kappa = 0.35, \lambda_g = 0.1$). Case 2: ($n = 50, c = 20, b = 15, \kappa = 0.5, \lambda_g = 0.1$). (a) Per node throughput T vs. η. (b) Expected E2E delay $\overline{D}_{e2e}$ vs. η. (c) Throughput capacity T_c vs η.

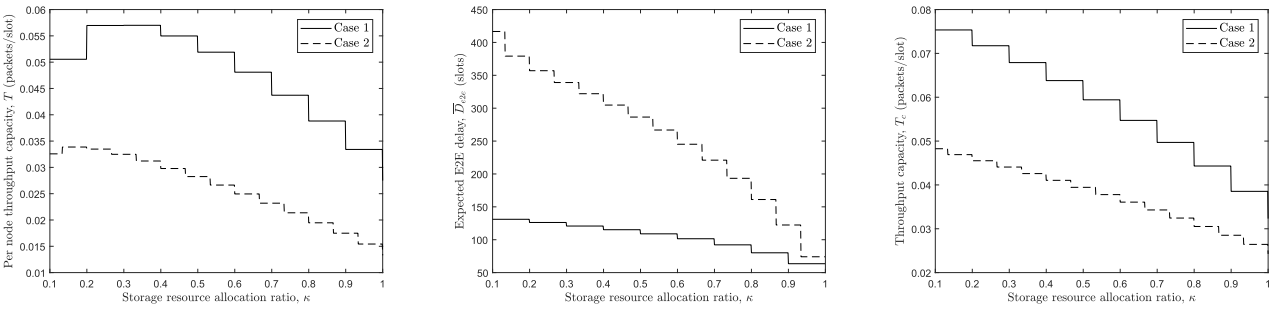


FIGURE 8. Effects of storage resource allocation on network performance. Case 1: ($n = 30, c = 15, b = 10, \eta = 0.3, \lambda_g = 0.1$). Case 2: ($n = 50, c = 20, b = 15, \eta = 0.6, \lambda_g = 0.1$). (a) Per node throughput T vs. κ. (b) Expected E2E delay $\overline{D}_{e2e}$ vs. κ. (c) Throughput capacity T_c vs κ.

cases considered here, indicating that our analytical framework is also highly efficient to evaluate the delay performance of a MANET. We can observe from Fig. 6 that as λ_g increases, $\overline{D}_{e2e}$ first increases somewhat and then decreases monotonically. Such behavior is due to the reason that the effects of λ_g on $\overline{D}_{e2e}$ are two folds. On one hand, a larger λ_g leads to a more congested MANET with a larger source delay $\overline{D}_s$ and a larger relay delay $\overline{D}_r$; on the other hand, a more congested MANET indicates that a packet is more likely to be delivered through a direct S-D operation, thus a smaller ratio of $\overline{D}_r$ will compose $\overline{D}_{e2e}$ (can be seen in (29)). As λ_g increases, these two effects become dominant alternatively, causing the increase-decrease phenomena of $\overline{D}_{e2e}$.

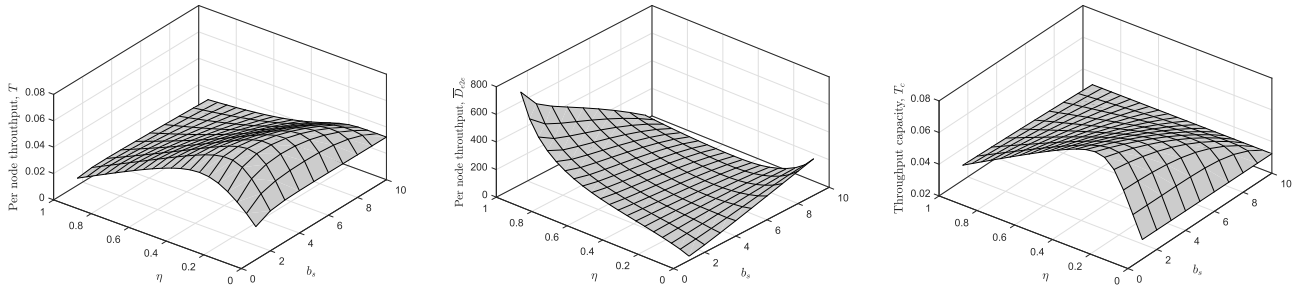
C. PERFORMANCE DISCUSSIONS

Based on our theoretical performance evaluation, we present numerical results to illustrate the impacts of resource allocation on network performance.

In order to highlight the effects of transmission resource allocation on network performance, we draw Fig. 7 to show how the per node throughput, expected E2E delay and throughput capacity varies with the transmission resource allocation ratio η, where the storage resource allocation ratio κ is fixed, and two cases, i.e., Case 1 ($n = 30, c = 15, b = 10, \kappa = 0.35, \lambda_g = 0.1$) and Case 2 ($n = 50, c = 20, b = 15, \kappa = 0.5, \lambda_g = 0.1$), are considered. Fig. 7(a) shows that the curve of throughput varying with η is unimodal under both the two cases. As η increases, T first increases to its maximum and then decreases. Fig. 7(b) shows that the behaviors of $\overline{D}_{e2e}$ varying with η can be different under different network settings. As η increases, $\overline{D}_{e2e}$ under Case 1 monotonically increases, while $\overline{D}_{e2e}$ under Case 2 decreases first and then increases. Fig. 7(c) shows that the behavior of T_c is similar to that of T in Fig. 7(a). We can observe from Fig. 7 that the transmission resource allocation does impose great effects on network performance.

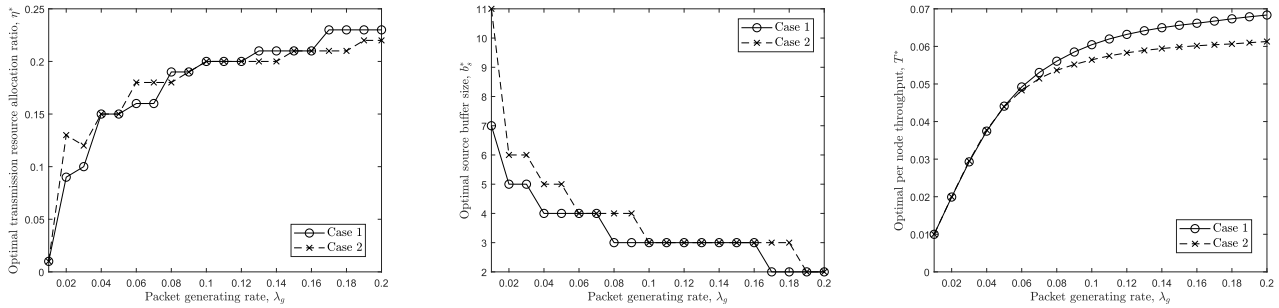
We then summarize in Fig. 8 that how network performance varies with the storage resource allocation ratio κ, where the transmission resource allocation ratio η is fixed and two cases, i.e., Case 1 ($n = 30, c = 15, b = 10, \eta = 0.3, \lambda_g = 0.1$) and Case 2 ($n = 50, c = 20, b = 15, \eta = 0.6, \lambda_g = 0.1$), are considered. As κ increases, Fig. 8(a) shows that T first increases and then decreases, while Fig. 8(b) and Fig. 8(c) show that $\overline{D}_{e2e}$ and T_c decrease monotonically. We can observe from Fig. 8 that the storage resource allocation also imposes great effects on network performance.

We further plot Fig. 9 to illustrate how the network performance varies with both the storage resource allocation and transmission resource allocation, under the network setting of ($n = 30, c = 15, b = 10, \lambda_g = 0.1$). Fig. 9(a) shows the curve of T varying with η (resp. b_s) is unimodal



(a) (b) (c)

FIGURE 9. Effects of resource allocation on network performance. Network setting: ($n = 30, c = 15, b = 10, \lambda_g = 0.1$). (a) Per node throughput T. (b) Expected E2E delay $\overline{D}_{e2e}$. (c) Throughput capacity T_c.

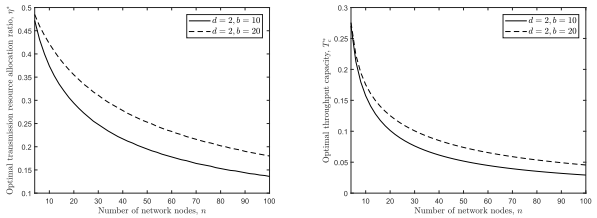


(a) (b) (c)

FIGURE 10. Optimal resource allocation for maximizing per node throughput. Case 1: ($n = 30, c = 15, b = 10$); Case 2: ($n = 50, c = 20, b = 15$). (a) Optimal transmission resource allocation ratio η^*. (b) Optimal source buffer size b_s^*. (c) Optimal per node throughput T^*.

under different settings of b_s (resp. η). The behavior of $\overline{D}_{e2e}$ in Fig. 9(b) is complicated. For example, with a small η, $\overline{D}_{e2e}$ increases as b_s increases, while with a large η, $\overline{D}_{e2e}$ decreases as b_s increases. Fig. 9(c) shows that T_c always decreases as b_s increases, which agrees with the analysis in Section VI-C. From Fig. 9, we have the following important observations: i) appropriate resource allocation can bring the benefit of performance enhancement for MANETs; ii) the optimal resource allocation configuration is different for different performance objectives (i.e., per node throughput, expected E2E delay and throughput capacity); iii) usually, in order to achieve good network performance, we need a small η and a small κ, i.e., a network node needs to be selfless to allocate more storage resource and more transmission resource for storing and forwarding packets of other nodes. These observations can serve as useful guidelines for the practical configurations and operations of MANETs.

Next, we summarize in Fig. 10 how the optimal resource allocation configuration (η^*, κ^*) for maximizing the per node throughput varying with the packet generating rate λ_g, where Case 1 ($n = 30, c = 15, b = 10$) and Case 2 ($n = 50$, $c = 20, b = 15$) are considered. We can see from Fig. 10 that as λ_g increases, the general trend of η^* is rising, while that of b_s^* is falling. It indicates that for achieving the maximal per node throughput, as the network traffic becomes heavy we need to increase the ratio of S-R operation while decreasing the ratio of storage space allocated to the source buffer. Comparing Fig. 10(c) with Fig. 5, we can see again that the optimal resource allocation can lead to distinct performance enhancement of MANETs, especially for Case 2. As mentioned before, the procedures of determining optimal resource allocation for minimizing expected E2E delay is similar to that for maximizing per node throughput, we skip over presenting the figure here, and readers can apply our simulator to explore the issues of interest freely.



(a) (b)

FIGURE 11. Optimal resource allocation for maximizing throughput capacity. Case 1: ($d = 2, b = 10$). Case 2: ($d = 2, b = 20$). (a) η^* vs. n. (b) T_c^* vs. n.

Finally, we summarize in Fig. 11 how the optimal transmission resource allocation ratio η^* varies with the number of network nodes n for the throughput capacity maximization (here the optimal storage resource allocation is always $b_s^* = b - 1$). Two cases, i.e., Case 1 ($d = 2, b = 10$) and Case 2 ($d = 2, b = 20$), are considered. Fig. 11(a) shows that η^* monotonically decreases as n^* increases, and it never

exceeds 0.5 which is in accordance with the analysis in Section VI-C. It indicates that as the network size becomes large, each network node needs to be more selfless, i.e., allocating more transmission resource for forwarding packets of other nodes. We can also see that η^* with a larger storage resource is always higher than that with a smaller one. Fig. 11(b) shows that T_c^* monotonically decreases as n increases, and a larger storage resource can achieve a higher T_c^*.

VIII. CONCLUSION AND FUTURE WORKS

The resource allocation for performance enhancement of MANETs was investigated in this paper. For a general resource allocation configuration, we first established an efficient analytical framework for modeling the network dynamical behaviors and deriving the exact expressions of per node throughput, expected E2E delay and throughput capacity. With the help of the performance evaluation results, we further proposed efficient algorithms to determine the optimal resource allocation configuration.

Some interesting conclusions can be drawn from this study: i) Appropriate resource allocation can bring the benefit of performance enhancement for MANETs; ii) The optimal resource allocation configuration is different for different performance objectives; iii) Usually, in order to achieve good network performance, a network node needs to be selfless to allocate more storage resource and more transmission resource for storing and forwarding packets of other nodes; iv) As the network size becomes large, in order to gain performance enhancement, each network node needs to allocate more resource for forwarding packets of other nodes. These findings can serve as useful guidelines for the practical configurations and operations of MANETs.

Notice that the analysis framework and resource allocation methods developed in this paper is based on the "uniform type" mobility model, so one of our future research directions is to explore the performance modeling and enhancement under "non-uniform type" mobility model and in static ad hoc networks. Moreover, in this study, we consider the unicast traffic pattern and the number of traffic flows m is the same as the number of nodes n. For the case of $m < n$, there will be a part of network nodes serving as pure relay nodes, i.e., there is no traffic originated from these nodes. Therefore, the resource allocation issue in MANETs under other traffic patterns (i.e., unicast with $m < n$, multicast) is another appealing direction to be investigated.

REFERENCES

[1] R. Ramanathan and J. Redi, "A brief overview of ad hoc networks: Challenges and directions," *IEEE Commun. Mag.*, vol. 40, no. 5, pp. 20–22, May 2002.
[2] M. N. Tehrani, M. Uysal, and H. Yanikomeroglu, "Device-to-device communication in 5G cellular networks: Challenges, solutions, and future directions," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 86–92, May 2014.
[3] H. Shariatmadari, R. Ratasuk, S. Iraji, A. Laya, T. Taleb, R. Jäntti, and A. Ghosh, "Machine-type communications: Current status and future perspectives toward 5G systems," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 10–17, Sep. 2015.
[4] J. Andrews, S. Shakkottai, R. Heath, N. Jindal, M. Haenggi, R. Berry, D. Guo, M. Neely, S. Weber, S. Jafar, and A. Yener, "Rethinking information theory for mobile ad hoc networks," *IEEE Commun. Mag.*, vol. 46, no. 12, pp. 94–101, Dec. 2008.
[5] A. Goldsmith, M. Effros, R. Koetter, M. Medard, A. Ozdaglar, and L. Zheng, "Beyond Shannon: The quest for fundamental performance limits of wireless ad hoc networks," *IEEE Commun. Mag.*, vol. 49, no. 5, pp. 195–205, May 2011.
[6] M. Grossglauser and D. N. C. Tse, "Mobility increases the capacity of ad hoc wireless networks," *IEEE/ACM Trans. Netw.*, vol. 10, no. 4, pp. 477–486, Aug. 2002.
[7] M. J. Neely and E. Modiano, "Capacity and delay tradeoffs for ad hoc mobile networks," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 1917–1937, Jun. 2005.
[8] A. El Gamal, J. Mammen, B. Prabhakar, and D. Shah, "Optimal throughput-delay scaling in wireless networks—Part I: The fluid model," *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2568–2592, Jun. 2006.
[9] G. Sharma, R. Mazumdar, and N. B. Shroff, "Delay and capacity tradeoffs in mobile ad hoc networks: A global perspective," *IEEE/ACM Trans. Netw.*, vol. 15, no. 5, pp. 981–992, Oct. 2007.
[10] X. Wang, W. Huang, S. Wang, J. Zhang, and C. Hu, "Delay and capacity tradeoff analysis for motioncast," *IEEE/ACM Trans. Netw.*, vol. 19, no. 5, pp. 1354–1367, Oct. 2011.
[11] Y. Wang, X. Chu, X. Wang, and Y. Cheng, "Optimal multicast capacity and delay tradeoffs in MANETs: A global perspective," in *Proc. IEEE INFOCOM*, Apr. 2011, pp. 640–648.
[12] Y. Qin, X. Tian, W. Wu, and X. Wang, "Mobility weakens the distinction between multicast and unicast," *IEEE/ACM Trans. Netw.*, vol. 24, no. 3, pp. 1350–1363, Jun. 2016.
[13] J. Liu, M. Sheng, Y. Xu, J. Li, and X. Jiang, "End-to-end delay modeling in buffer-limited MANETs: A general theoretical framework," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 498–511, Jan. 2016.
[14] J. Liu, Y. Xu, Y. Shen, X. Jiang, and T. Taleb, "On performance modeling for MANETs under general limited buffer constraint," *IEEE Trans. Veh. Technol.*, vol. 66, no. 10, pp. 9483–9497, Oct. 2017.
[15] Y. Fang, Y. Zhou, X. Jiang, and Y. Zhang, "Practical performance of MANETs under limited buffer and packet lifetime," *IEEE Syst. J.*, vol. 11, no. 2, pp. 995–1005, Jun. 2017.
[16] K. Lee, Y. Yi, J. Jeong, H. Won, I. Rhee, and S. Chong, "Max-contribution: On optimal resource allocation in delay tolerant networks," in *Proc. IEEE INFOCOM*, Mar. 2010, pp. 1–9.
[17] X. Sun and S. Wang, "Resource allocation scheme for energy saving in heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 8, pp. 4407–4416, Aug. 2015.
[18] Q. Wu, M. Tao, D. W. K. Ng, W. Chen, and R. Schober, "Energy-efficient resource allocation for wireless powered communication networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 2312–2327, Mar. 2016.
[19] C. Wang, C. Liang, F. R. Yu, Q. Chen, and L. Tang, "Computation offloading and resource allocation in wireless cellular networks with mobile edge computing," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 4924–4938, Aug. 2017.
[20] L. Liang, G. Y. Li, and W. Xu, "Resource allocation for D2D-enabled vehicular communications," *IEEE Trans. Commun.*, vol. 65, no. 7, pp. 3186–3197, Jul. 2017. doi: 10.1109/TCOMM.2017.2699194.
[21] A. Ahmad, S. Ahmad, M. H. Rehmani, and N. Ul Hassan, "A survey on radio resource allocation in cognitive radio sensor networks," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 888–917, 2nd Quart., 2015.
[22] M. E. Tanab and W. Hamouda, "Resource allocation for underlay cognitive radio networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 1249–1276, 2nd Quart., 2017.
[23] S. Giannecchini, M. Caccamo, and C.-S. Shih, "Collaborative resource allocation in wireless sensor networks," in *Proc. 16th Euromicro Conf. Real-Time Syst.*, Catania, Italy, Jul. 2004, pp. 35–44.
[24] J. Yuan and W. Yu, "Joint source coding, routing and power allocation in wireless sensor networks," *IEEE Trans. Commun.*, vol. 56, no. 6, pp. 886–896, Jun. 2008.
[25] X. Wang, D. Wang, H. Zhuang, and S. D. Morgera, "Fair energy-efficient resource allocation in wireless sensor networks over fading TDMA channels," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 7, pp. 1063–1072, Sep. 2010.
[26] J. Liu, K. Xiong, P. Fan, and Z. Zhong, "Resource allocation in wireless powered sensor networks with circuit energy consumption constraints," *IEEE Access*, vol. 5, pp. 22775–22782, 2017.

[27] W. Lu, T. Nan, Y. Gong, M. Qin, X. Liu, Z. Xu, and Z. Na, "Joint resource allocation for wireless energy harvesting enabled cognitive sensor networks," *IEEE Access*, vol. 6, pp. 22480–22488, 2018.
[28] P. Nain, D. Towsley, B. Liu, and Z. Liu, "Properties of random direction models," in *Proc. IEEE INFOCOM*, Mar. 2005, pp. 1897–1907.
[29] T. G. Robertazzi, *Computer Networks and Systems: Queueing Theory and Performance Evaluation*. New York, NY, USA: Springer, 2012.
[30] H. Daduna, *Queueing Networks With Discrete Time Scale: Explicit Expressions for the Steady State Behavior of Discrete Time Stochastic Networks*. Berlin, Germany: Springer-Verlag, 2001.
[31] U. Niesen, D. Shah, and G. W. Wornell, "Adaptive alternating minimization algorithms," *IEEE Trans. Inf. Theory*, vol. 55, no. 3, pp. 1423–1429, Mar. 2009.
[32] A. Granas and J. Dugundji, *Fixed Point Theory and Applications*. New York, NY, USA: Springer-Verlag, 2003.
[33] L. De Haan and A. Ferreira, *Extreme Value Theory: An Introduction*. New York, NY, USA: Springer, 2007.
[34] J. Liu and Y. Xu. (2019). *Simulation Codes for Exploring Resource Allocation in MANETs*. [Online]. Available: https://www.researchgate.net/publication/331824181

JIA LIU (S'10–M'12) received the Ph.D. degree from the School of Systems Information Science, Future University Hakodate, Japan, in 2016. He is currently an Assistant Professor with the Center for Cybersecurity Research and Development, National Institute of Informatics, Japan. His research interests include mobile ad hoc networks, 5G communication systems, D2D communications, the IoT, physical layer security, and cyber security. He has published around 30 technical papers at premium international journals and conferences, including the IEEE TWC, the IEEE TVT, the IEEE INFOCOM, and *Computer Networks*. He has received the 2016 IEEE Sapporo Section Encouragement Award and the Best Paper Award from NaNA 2017.

YANG XU (S'09–M'12) received the B.E. degree from the School of Telecommunications Engineering and the Ph.D. degree from the Department of Communication and Information Systems, Xidian University, Xi'an, China, in 2006 and 2014, respectively, where she is currently a Lecturer with the School of Economics and Management, The State Key Laboratory of Integrated Services Network. She has published about 30 papers at premium international journals, including the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATION, and *Computer Networks*. She has published in conferences including the IEEE ICC and the IEEE WCNC. Her current research interests include physical layer security, block-chain, routing protocol design, and network performance evaluation.

ZHAO LI (S'08–M'10) received the B.S. degree in telecommunications engineering and the M.S. and Ph.D. degrees in communication and information systems from Xidian University, Xi'an, China, in 2003, 2006, and 2010, respectively, where he is currently an Associate Professor with the School of Cyber Engineering. He was a Visiting Scholar and then a Research Scientist with the Real-Time Computing Laboratory, Department of Electrical Engineering and Computer Science, University of Michigan, from 2013 to 2015. He has published over 40 technical papers at premium international journals and conferences, including *Computer Communications*, the IEEE TWC, and the IEEE INFOCOM. His research interests include wireless communications, 5G communication systems, resource allocation, interference management, the IoT, and physical layer security.