Quality-of-Service Differentiation in Buffer-aided Cooperative Free Space Optical Communication Systems

Chadi Abou-Rjeily, *Senior Member IEEE*, and Wissam Fawaz, *Senior Member IEEE*

***Abstract*—In this paper, the problem of Quality-of-Service (QoS) differentiation is studied in the context of Buffer-Aided (BA) cooperative Free Space Optical (FSO) communication systems. This is particularly true since the existing relevant literature overlooked the possibility of preferential treatment of data packets based on their delay requirements. Inspired by this observation, this paper proposes to classify the data packets emanating from the source node into either Delay Tolerant (DT) packets or Non Delay Tolerant (NDT) ones and to service these packets according to this classification at the relay node. Priority queueing is first introduced in the relay's buffer with a view to improving the delay experienced by NDT packets. Class of service mutation is then proposed as a starvation mitigation strategy to better manage the interesting dynamics resulting from the co-existence of packets having different QoS requirements in the same buffer. The various performance measures of interest for the explored QoS-aware communication system are both evaluated mathematically based on a Markov chain analysis and validated through extensive simulations. An asymptotic analysis is also carried out highlighting the dependence of the performance on the system parameters in an intuitive manner.**

***Index Terms*—Free space optics, relaying, buffer, priority queuing, delay tolerance, asymptotic analysis, quality of service, class-of-service mutation.**

I. Introduction

Today's Internet is a melting pot for a plethora of applications having different Quality-of-Service (QoS) requirements. One critical QoS performance measure is the delay where the delay requirements vary from one application to another depending on the nature and the type of the considered application [1]. For example, an application involving an online virtual interactive environment is expected to have more stringent delay expectations than a database backup application or even a non-real time file transfer application. In light of this discussion, this work, unlike previous studies, proposes to broadly categorize the data packets travelling through Buffer-aided (BA) cooperative Free Space Optical (FSO) communication systems into Delay Tolerant (DT) and Non Delay Tolerant (NDT) packets. More importantly, this study explores different scheduling, departure, and priority management procedures at the relay node for the purpose of better catering to the diversified delay requirements of the data packets traversing a BA cooperative FSO communication system.

The authors are with the Department of Electrical and Computer Engineering of the Lebanese American University (LAU), Byblos, Lebanon. (e-mails: {chadi.abourjeily,wissam.fawaz}@lau.edu.lb). This work is supported by the National Council for Scientific Research - Lebanon (CNRS-L), project #851.

The FSO communication technology holds the promise of offloading the overly crowded Radio Frequency (RF) spectrum by shifting data towards the optical spectrum [2]. Cooperative techniques were extensively studied in the context of FSO communication systems and took the form of incorporating a number of relays between the source and destination nodes so as to mitigate mainly the effect of the distance-dependent atmospheric-induced fading [3]–[9]. In this way, cooperative FSO communication allows the transmission of data packets originating from a source node to a destination node over the air through other communication nodes, designated as relays. It is important to note that the cooperative communication research landscape [3]–[9] has been traditionally predominated by studies assuming buffer-free relay-assisted communication. However, many later studies vouched for the great merit associated with BA relaying. This explains the recent spike of interest in BA relaying solutions in the context of RF systems [12]–[16], hybrid FSO/RF systems [17]–[19] as well as FSO systems [21]. The main idea behind BA relaying lies in equipping relays with buffers (or data packet queues) with the aim of storing information packets in these buffers until the quality of the link connecting the relay to destination becomes favourable for data packet transfer. This eventually has the advantage of improving the overall throughput of the system relative to the buffer-free case.

In [12]–[16], BA parallel relaying was studied in the context of RF systems. The *max-link* protocol was introduced in [12] where a time slot is devoted to either source (S) to relay (R) transmission or relay to destination (D) transmission. To enhance the system availability, communication under the *max-link* protocol takes place along the link having the best quality among all available S-R and R-D links. The *max-link* protocol, which was initially designed around Decode-and-Forward (DF) cooperation [12], was then extended to the case of Amplify-and-Forward (AF) cooperation in [13]. Improved versions of the *max-link* protocol were provided in [14]–[16]. Specifically, the authors in [14] proposed to improve the average packet delay by prioritizing the selection of the R-D links, in a bid to ensure a faster draining of the relays' buffers. Further improvements were introduced to the *max-link* protocol in [15], [16] by accounting for both buffer state information (BSI) and delay state information (DSI) in the relay selection procedure. In point of fact, the initial version of the *max-link* protocol based relay selection solely on channel state information (CSI). The work in [15] proposed to supplement the CSI with BSI in relay selection by

distinguishing between buffers that are full, empty, or neither full nor empty. On top of CSI and BSI, the authors in [16] considered also DSI by ensuring that packets exceeding a well-defined waiting time threshold are dropped from the relays' buffers.

In [17]–[21], BA parallel relaying was considered in the context of FSO and hybrid FSO/RF systems. Particularly, a link allocation strategy was proposed in [17] for a multiuser hybrid RF and mixed FSO/RF BA relay network. Therein, multiple RF mobile users were assumed to transmit data packets over a RF link to a DF relay, which in turn forwards the received packets through a FSO link that is supported by a RF backup system to the destination. The work in [17] was then refined in [18], where an efficient mixed RF and hybrid FSO/RF network that fully leverages the high transmission rates of multiuser scenarios was presented. BA relay selection was considered in [19] in the presence of both multiple relays equipped with infinite size queues as well as hybrid FSO/RF links. The authors of [20] proposed an adaptive transmission scheme for improving the statistical delay-throughput trade-off in the context of cooperative hybrid RF/FSO backhaul networks. The transmission scheme was shown to enhance the maximum supportable arrival rate of such networks. Finally, the problem of BA relay selection for cooperative FSO systems with multiple relays was investigated in [21] under the realistic assumption of finite size relays' buffers. The authors of [21] proposed multiple relaying protocols whose respective performance was contrasted and analyzed through a Markov chain analysis.

The process of routing delay tolerant (DT) as well as non-delay tolerant (NDT) data messages received wide attention in the context of different types of systems. For instance, the authors of [22] tackled the problem of scheduling the forwarding of real-time and non-real time data packets at sensor nodes with a view to reducing sensors energy consumptions and end-to-end data transmission delays. Their simulation results highlighted the ability of their so-called dynamic multilevel priority packet scheduling schemes to both reduce the average data waiting time and balance energy consumption. In [23], the concept of prioritized multi-stream traffic is studied to highlight the impact of prioritized uplink transmission on the performance of Internet of Things devices. In the same spirit, [24] surveys the different routing and data dissemination techniques that can be used when forwarding messages in delay tolerant networks. This study complements these and the many other existing studies by considering the problem of routing DT and NDT packets while accounting for the particularities of cooperative FSO communications.

To the authors' best knowledge, none of the surveyed studies revolving around BA cooperative communication considered the fundamental problem of QoS-aware data packet processing at the relay node. Therefore, this paper proposes to render BA cooperative FSO communication systems QoS-enabled by employing a priority-aware scheduling discipline at the relay node for the purpose of servicing the NDT and DT packets joining the relay's buffer in a way that is consistent with their respective delay requirements. In fact, to date, it has been assumed that the departure of data packets from the relay's buffer is scheduled according to the QoS-agnostic First Come First Serve (FCFS) discipline. However, FCFS is not well suited for dealing with prioritized data packets. This paper proposes thus to equip the relay with a priority queue that stores data packets according to their priorities while supporting departure of packets in order of priority as well. In this manner, NDT packets are stored in the relays' buffers in front of the DT packets to ensure an expedited clearance of NDT packets as compared to the DT ones. This results in reduced queueing delay values for NDT packets relative to the DT ones.

In this paper, the performance of the proposed QoS-aware system architecture is studied mathematically through a Markov chain analysis that derives closed form expressions for the delay as well as the packet loss experienced by the NDT and DT packets for any buffer size L. Valuable insights are also provided under the high Signal-to-Noise-Ratio (SNR) regime. In this case, asymptotic expressions are presented for the delay and packet loss metrics. The asymptotic analysis shows mainly that the performance of the system depends strongly on whether the relay is closer to the destination or closer to the source. Finally, this paper proposes the idea of class of service mutation in an attempt to strike a proper balance between the performance of NDT and DT packets. The idea of class of service mutation is driven mainly by the need to prevent NDT packets from severely penalizing the DT packets under low to medium SNR regimes. Under class of service mutation, an incoming DT packet is probabilistically treated as an NDT packet and consequently, is given equal status to an authentic NDT packet residing in the relay's buffer. This allows for a probabilistic improvement of the performance of DT packets without drastically degrading the performance of NDT packets. The main contributions of this paper can be summarized as follows:

- A novel priority-aware FSO BA relaying scheme with QoS differentiation is proposed. Unlike previous studies on BA cooperative FSO communication systems that considered First Come First Serve (FCFS) queues at the relay node without regard to quality of service differentiation, the present study demonstrates the ability of priority queueing at the relay node to ensure a preferential treatment of data packets based on their delay requirements.
- This paper introduces the novel concept of packet mutation whereby the delay of DT packets can be improved without incurring a violation of the delay requirements associated with the NDT packets.
- A Markov chain framework is developed with a view to accurately derive the packet loss and average packet delay for the considered BA cooperative FSO communication system. An asymptotic analysis is also provided for the sake of offering clear and intuitive insights on the system performance for high SNR.
- Lastly, an in-house discrete event simulator is used to validate the accuracy of the results emanating from our analytical Markov chain framework.

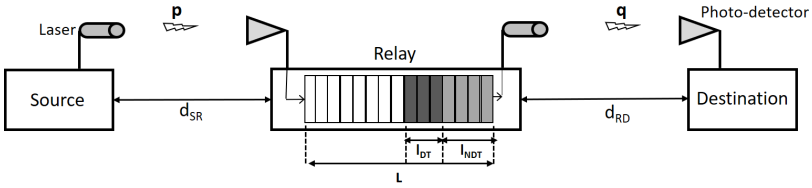

Fig. 1. System model of a BA FSO communication system with QoS differentiation.

II. System Model

Consider the case of a source (S) communicating with a destination (D) through a relay (R) equipped with a buffer of finite size denoted by L, as depicted in Fig. 1. Decode-and-Forward (DF) relaying is assumed where the packet received at R is decoded, stored, then retransmitted when the channel conditions along the R-D link are favorable.

The particularities of FSO transmissions render the relaying problem different from that considered in radio-frequency (RF) systems. (i): Unlike RF antennas that can be used for both transmission and reception, FSO nodes must be equipped with lasers for transmission and with photo-detectors for reception as shown in Fig. 1. These distinct components can be controlled independently, thus, offering additional degrees of freedom in the system design compared to RF communications. In particular, FSO nodes operate naturally in the full-duplex (FD) mode where simultaneous reception and transmission can take place at the photo-detector and laser placed at R, respectively. (ii): The different FSO links do not interfere with each other owing to the high directivity of the laser light beams. As such, multiple transmissions can take place simultaneously from S and R without interference. (iii): Unlike the broadcast nature of RF transmissions where the signal transmitted from a node can be overheard by all neighboring nodes, FSO links are LOS implying that a signal transmitted to a certain node cannot be detected by other nodes in the network. For example, in Fig. 1, the optical signal transmitted from S to R cannot reach D. The full-duplexity feature (in the absence of interference) has a direct impact on the queue dynamics where a packet can exit the buffer while another packet can concurrently enter this buffer. This radically affects the Markov chain analysis as well as the packet loss and delay derivations presented in this paper. It is worth noting that RF systems can operate in the FD mode where two antennas are deployed at the relay (one for reception and the other for transmission). However, this requires the implementation of involved self-loop-interference cancelation strategies unlike the case of FSO systems where the FD operation is spontaneous. Finally, the analysis presented in this paper can be readily applied to the case of RF FD communications under the assumptions of no self-loop-interference and absence of a direct link between S and D.

We assume that a packet is generated at S every time slot. With probability p_{NDT}, the packet is a high priority NDT packet that needs to be delivered to D with the least possible delay. With probability $p_{DT} = 1 - p_{NDT}$, the generated packet is DT with a lower priority relative to NDT packets. To minimize the queuing delays experienced by the NDT packets, these packets are stored in the relay's queue according to a head of the line priority discipline. This means that the NDT packets queue up at R in front of all of the DT packets that can start exiting the queue only when all the preceding NDT packets are transmitted to D. Given that the queue has a finite capacity, packet loss may be experienced at R where some packets might be dropped from the rear end of the queue. In this context, the NDT packets will also be given a higher priority in accessing the queue where a DT packet can be dropped from the rear end of a full queue to make room for an incoming higher priority NDT packet.

A Markov chain analysis will be adopted for studying the BA system [12]. A state of the Markov chain is represented by the numbers of NDT and DT packets present in the buffer. The state will be denoted by $\mathbf{l} \triangleq (l_{DT}, l_{NDT})$ where l_{DT} and l_{NDT} stand for the numbers of DT and NDT packets, respectively, with $0 \leq l_{tot} \triangleq l_{DT} + l_{NDT} \leq L$ resulting in $\frac{(L+1)(L+2)}{2}$ possible states. Defining the set $\mathcal{L}$ as $\mathcal{L} \triangleq \{(l_1, l_2) \mid l_1 \geq 0\ ,\ l_2 \geq 0\ ,\ 0 \leq l_1 + l_2 \leq L\}$, the evolution between the states will be described by the transition probabilities $\{t_{\mathbf{l},\mathbf{l}'}\}_{(\mathbf{l},\mathbf{l}')\in\mathcal{L}^2}$ where $t_{\mathbf{l},\mathbf{l}'}$ stands for the probability of moving from state $\mathbf{l}$ to state $\mathbf{l}'$.

The communications between the nodes will be established through intensity-modulated with direct-detection (IM/DD) FSO links where the transmitted symbols are carved from a binary On-Off-Keying (OOK) signal set. We consider the case of background noise limited receivers where the shot noise caused by background radiation is dominant with respect to the other noise components [3]. This results in an IM/DD system corrupted by a signal-independent additive white Gaussian noise (AWGN) whose variance will be denoted by N_0. We will also adopt the widely approved gamma-gamma turbulence-induced fading model for capturing the scintillation along the FSO links. An FSO link will be in outage if the signal-to-noise ratio (SNR) falls below a threshold level γ_{th} that ensures the signal decodability. For gamma-gamma fading with AWGN noise, the outage probability of a link of length d can be calculated from [8]:

$$P_{\text{out}}(d, P_M) = \frac{1}{\Gamma(\alpha(d))\Gamma(\beta(d))} G_{1,3}^{2,1}\left[\frac{\alpha(d)\beta(d)}{G(d)[P_M/N_{\text{link}}]}\middle|\begin{matrix}1\\ \alpha(d), \beta(d), 0\end{matrix}\right], \quad (1)$$

where $G_{p,q}^{m,n}[.]$ is the Meijer G-function and $\Gamma(.)$ is the gamma function. $G(d) = \left(\frac{d_{\text{SD}}}{d}\right)^2 e^{-\sigma(d-d_{\text{SD}})}$ is the gain that arises when the links are shorter that the S-D link of length d_{SD},

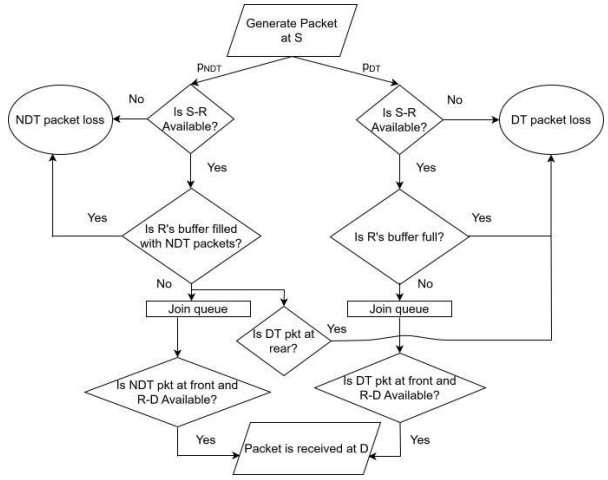

Fig. 2. Flow chart of the priority-aware BA relaying protocol.

where σ is the attenuation coefficient [3]. The parameters of the gamma-gamma distribution are given by:

$$\alpha(d) = \left[\exp\left(0.49\sigma_R^2(d)/(1+1.11\sigma_R^{12/5}(d))^{7/6}\right) - 1\right]^{-1}, \tag{2}$$

$$\beta(d) = \left[\exp\left(0.51\sigma_R^2(d)/(1+0.69\sigma_R^{12/5}(d))^{5/6}\right) - 1\right]^{-1}, \tag{3}$$

where the distance-dependent Rytov variance is given by $\sigma_R^2(d) = 1.23C_n^2k^{7/6}d^{11/6}$ with k and C_n^2 denoting the wave number and refractive index structure parameter, respectively. In (1), P_M stands for the optical power margin that is normalized by N_{link}, which stands for the total number of links. For a single-relay system, N_{link} is equal to 2, following from evenly splitting the optical power along the S-R and R-D links in the absence of channel state information. The power margin is related to the threshold SNR by $P_M = \eta\sqrt{\frac{E_b}{N_0\gamma_{th}}}$, where η is the optical-to-electrical conversion ratio and E_b is the signal energy per bit.

Following from (1), the outage probabilities along the S-R and R-D links will be denoted by:

$$p = P_{\text{out}}(d_{\text{SR}}, P_M) \;\; ; \;\; q = P_{\text{out}}(d_{\text{RD}}, P_M), \tag{4}$$

where d_{SR} and d_{RD} stand for the lengths of the S-R and R-D links, respectively.

III. Performance Analysis

A. BA Priority Relaying

A flow chart of the priority-aware relaying protocol is shown in Fig. 2. (i): If the buffer is empty, no packets can be transmitted from R. In this case, any incoming NDT or DT packet can enter the buffer. (ii): If the buffer is not empty, the sequence of actions taken by R is as follows:

- At the beginning of each time slot, R attempts to send a packet to D where a successful attempt takes place with probability $1-q$. The transmitted packet is of the NDT type if $l_{NDT} \neq 0$ and of the DT type otherwise if $l_{DT} \neq 0$.
- R then processes an incoming packet received from S (if any). If the buffer is not full (after the transmission attempt), any incoming packet can be stored in the buffer, where a NDT (resp. DT) packet is stored behind all of the already present NDT (resp. DT) packets. If the buffer is full and a DT packet is received at R, then this low priority packet will be dropped. If a NDT packet is received, R drops any DT packet already stored in the queue (if any) in order to accommodate the incoming high priority NDT packet. If all of the already stored packets are of the NDT type, then the received NDT packet will be dropped.

Based on the above proposed relay buffering strategy, the NDT packets are given preference whether in exiting or in entering the buffer at R, thus reducing their corresponding queuing delays and packet losses. NDT and DT packets arrive at R with the following probabilities:

$$\alpha_{NDT} = (1-p)p_{NDT} \;\; ; \;\; \alpha_{DT} = (1-p)p_{DT}, \tag{5}$$

where these packets can be detected at R only if the S-R link is not in outage. With probability $1-\alpha_{NDT}-\alpha_{DT} = p$, no packets arrive at R following from the outage of the S-R link.

B. Transition Probabilities and Steady-State Distribution

We will next evaluate the transition probabilities $\{t_{\mathbf{l},\mathbf{l}'}\}_{(\mathbf{l},\mathbf{l}')\in\mathcal{L}^2}$. Following from the full-duplexity of FSO communications where R can simultaneously transmit and receive within the same time slot, the transition probabilities can be written under the following general form in case the buffer is not empty (so that a packet can be transmitted):

$$t_{\mathbf{l},\mathbf{l}'} = qt_{\mathbf{l},\mathbf{l}'}^{(0)} + (1-q)t_{\mathbf{l},\mathbf{l}'}^{(1)} \;\; ; \;\; \mathbf{l} \neq (0,0), \tag{6}$$

where $t_{\mathbf{l},\mathbf{l}'}^{(0)}$ (resp. $t_{\mathbf{l},\mathbf{l}'}^{(1)}$) stands for the conditional transition probability when the R-D link is (resp. is not) in outage. In (6), the probability $(1-q)t_{\mathbf{l},\mathbf{l}'}^{(1)}$ accounts for the event where the FD relay concurrently transmits and receives. For example, $t_{\mathbf{l},\mathbf{l}'}^{(1)} = \alpha_{NDT}$ (resp. $t_{\mathbf{l},\mathbf{l}'}^{(1)} = \alpha_{DT}$) implies that a packet was transmitted from R and a NDT (resp. DT) packet was successfully received at R.

The state $\mathbf{l}'$ can be written as $\mathbf{l}' = \mathbf{l} + (\delta_1, \delta_2)$. and the following cases arise.

Case 1: $\mathbf{l} = (0,0)$. In this case, the buffer is empty and no packets can be transmitted from R to D. This results in $t_{(0,0),(0,0)} = 1-\alpha_{NDT}-\alpha_{DT}$, $t_{(0,0),(0,1)} = \alpha_{NDT}$ and $t_{(0,0),(1,0)} = \alpha_{DT}$ depending on whether no packet, a NDT packet or a DT packet is received at R.

Case 2: $\mathbf{l} = (0,L)$, where the buffer is full with NDT packets. In this case, if the R-D link is in outage (with probability q), the buffer remains full and no incoming packets can be accommodated at R since all packets in the buffer have the highest priority. Otherwise, a NDT packet exits the queue at the beginning of the time slot resulting in the following possibilities. (i): If a NDT packet arrives (with probability α_{NDT}), $(\delta_1, \delta_2) = (0,0)$ since one NDT packet is transmitted to D while another NDT packet is received from S. (ii): If a DT packet arrives (with probability α_{DT}), $(\delta_1, \delta_2) = (1,-1)$ since one NDT packet is transmitted to D while a DT packet

is received from S. (iii): Finally, if no packets arrive at R (with probability $1-\alpha_{NDT}-\alpha_{DT}$), $(\delta_1,\delta_2)=(0,-1)$. Therefore, $t_{(0,L),(0,L)}=q+(1-q)\alpha_{NDT}$, $t_{(0,L),(1,L-1)}=(1-q)\alpha_{DT}$ and $t_{(0,L),(0,L-1)}=(1-q)(1-\alpha_{NDT}-\alpha_{DT})$.

Case 3: $\mathbf{l}=(L,0)$, where the buffer is full with DT packets. In this case, the following transitions are possible. (i): $t_{(L,0),(L,0)}=q(1-\alpha_{NDT})+(1-q)\alpha_{DT}$, where the occupancy of the buffer remains the same either if the R-D link is in outage and no NDT packet arrives at R (otherwise this packet will take the place of a low priority DT packet in the queue) or if this link is not in outage and a DT packet arrives. In the last case, a DT packet is transmitted while another one is received which is possible since R is full-duplex. (ii): $t_{(L,0),(L-1,0)}=(1-q)(1-\alpha_{NDT}-\alpha_{DT})$, where a DT packet is transmitted while no packet is received. (iii): $t_{(L,0),(L-1,1)}=(1-q)\alpha_{NDT}+q\alpha_{NDT}=\alpha_{NDT}$ since with probability $(1-q)\alpha_{NDT}$ a DT packet will be transmitted and a NDT packet will be received, while with probability $q\alpha_{NDT}$ no packet is transmitted implying that the incoming NDT packet will take the place of a DT packet.

Case 4: $\mathbf{l}=(l_{DT},l_{NDT})$ with $l_{NDT}\neq 0$ and $l_{DT}+l_{NDT}\neq L$. The possible values of (δ_1,δ_2) are as follows. (i): $(\delta_1,\delta_2)=(0,0)$ with probability $q(1-\alpha_{NDT}-\alpha_{DT})+(1-q)\alpha_{NDT}$, where either no packet is transmitted and no packet is received or a NDT packet is transmitted and another one is received. (ii): $(\delta_1,\delta_2)=(0,1)$ (resp. $(\delta_1,\delta_2)=(1,0)$) with probability $q\alpha_{NDT}$ (resp. $q\alpha_{DT}$), where no packet is transmitted and a NDT (resp. DT) packet is received. (iii): $(\delta_1,\delta_2)=(0,-1)$ with probability $(1-q)(1-\alpha_{NDT}-\alpha_{DT})$, where a NDT packet is transmitted and no packet is received. (iv): $(\delta_1,\delta_2)=(1,-1)$ with probability $(1-q)\alpha_{DT}$, where a NDT packet is transmitted and a DT packet is concurrently received within the same time slot.

Case 5: $\mathbf{l}=(l_{DT},0)$ with $l_{DT}\neq 0$ and $l_{DT}\neq L$. Similar to case 4, the possible transitions and their corresponding probabilities are given by $t_{\mathbf{l},\mathbf{l}}=q(1-\alpha_{NDT}-\alpha_{DT})+(1-q)\alpha_{DT}$, $t_{\mathbf{l},\mathbf{l}+(0,1)}=q\alpha_{NDT}$, $t_{\mathbf{l},\mathbf{l}+(1,0)}=q\alpha_{DT}$, $t_{\mathbf{l},\mathbf{l}+(-1,0)}=(1-q)(1-\alpha_{NDT}-\alpha_{DT})$ and $t_{\mathbf{l},\mathbf{l}+(-1,1)}=(1-q)\alpha_{NDT}$.

Case 6: $\mathbf{l}=(l_{DT},l_{NDT})$ with $l_{DT}\neq 0$, $l_{NDT}\neq 0$ and $l_{DT}+l_{NDT}=L$. In this case, the packet to be transmitted is NDT while an incoming NDT packet will replace a DT packet in the queue. The scenarios that might arise are as follows. (i): $(\delta_1,\delta_2)=(0,0)$ with probability $(1-q)\alpha_{NDT}+q(1-\alpha_{NDT})$, where when no packet can be transmitted (with probability q), no NDT packet must arrive to keep the occupancy of the buffer unchanged since this packet will take the place of a DT packet in the queue. (ii): Similar to the previous cases $(\delta_1,\delta_2)=(0,-1)$ and $(\delta_1,\delta_2)=(1,-1)$ with probabilities $(1-q)(1-\alpha_{NDT}-\alpha_{DT})$ and $(1-q)\alpha_{DT}$, respectively. (iii): $(\delta_1,\delta_2)=(-1,1)$ with probability $q\alpha_{NDT}$, where, for a full queue, a DT packet needs to be dropped to accomodate an incoming higher priority NDT packet.

Putting the probabilities $\{t_{\mathbf{l},\mathbf{l}'}\}_{(\mathbf{l},\mathbf{l}')\in\mathcal{L}^2}$ together to form the state transition matrix $\mathbf{T}$ results in the following expression for the steady-state probability vector [12]:

$$\pi=(\mathbf{T}-\mathbf{I}+\mathbf{B})^{-1}\mathbf{b}, \tag{7}$$

where, denoting the number of states by $L_s\triangleq\frac{(L+1)(L+2)}{2}$, $\mathbf{I}$ is the $L_s\times L_s$ identity matrix, $\mathbf{B}$ is the $L_s\times L_s$ matrix whose elements are all equal to 1 and $\mathbf{b}$ is the $L_s\times 1$ vector whose elements are all equal to 1. The components of the $L_s\times 1$ vector π will be numbered as $\pi_{l_{DT},l_{NDT}}$, which stands for the probability of having l_{DT} DT packets and l_{NDT} NDT packets in the buffer at steady-state for $(l_{DT},l_{NDT})\in\mathcal{L}$.

A simple closed-form evaluation of the stationary distribution in (7) for an arbitrary value of L seems to be out of reach. An adequate remedy to this limitation resides in evaluating the steady-state marginal distributions $\{\pi_l^{(NDT)}\}_{l=0}^{L}$ and $\{\pi_l^{(tot)}\}_{l=0}^{L}$ for the number of NDT packets and the total number of packets, respectively. This approach not only offers more insights into the achievable performance levels, but is also sufficient for evaluating the packet-loss and packet-delay in exact closed-forms as will be highlighted in the subsequent section.

The evaluation of $\{\pi_l^{(NDT)}\}_{l=0}^{L}$ is possible since the NDT traffic is not affected by the DT traffic. In fact, the arrival of a DT packet will not affect the number of NDT packets present in the queue since the DT packets are queued behind the NDT packets (if empty buffering positions are available). Similarly, a DT packet can not be transmitted to D unless all NDT packets have been transmitted from the queue. On the other hand, observing the total number of packets without any distinction between their types is equivalent to analyzing a queue where packets arrive with probability $1-p$ and leave with probability $1-q$ which directly leads to the evaluation of $\{\pi_l^{(tot)}\}_{l=0}^{L}$ in a straightforward manner. It is worth highlighting that this approach does not hold for the DT packets since the arrival and departure of these packets is highly influenced by the number of NDT packets present in the queue. Moreover, even the arrival of a single NDT packet will have its impact on the number of DT packets in the queue.

Proposition *1:* Defining the probabilities $\{\mathfrak{p}_l(\lambda)\}_{l=0}^{L}$ as:

$$\begin{cases}\mathfrak{p}_0(\lambda)=q\left[\frac{1-r^{L+1}}{1-r}-(1-q)\right]^{-1}\\ \mathfrak{p}_l(\lambda)=\frac{1}{q}r^l\mathfrak{p}_0(\lambda)\ ,\ l=1,\ldots,L\end{cases};\ r\triangleq\frac{q\lambda}{(1-q)(1-\lambda)}, \tag{8}$$

then:

$$\pi_l^{(NDT)}=\mathfrak{p}_l(\alpha_{NDT})\ ;\ \pi_l^{(tot)}=\mathfrak{p}_l(1-p)\ \text{ for }\ l=0,\ldots,L. \tag{9}$$

Proof: The proof is provided in Appendix A. ■

C. Packet Loss (PL) and Average Packet Delay (APD)

First, we evaluate the Packet Loss (PL) and Average Packet Delay (APD) for the NDT packets. NDT packet loss can be experienced either at S or at R. At S, a NDT packet generated with probability p_{NDT} will not reach R if the S-R link is in outage with probability p. On the other hand, with probability α_{NDT}, the NDT packet will arrive at R. Now, if the buffer is not full or if it was full and a packet was successfully transmitted from the buffer at the beginning of the time slot, then the arriving NDT packet can be accommodated into the buffer. Similarly, if the buffer remains full (after the transmission attempt at the beginning of the time slot) and

there exists at least one DT packet in the queue, then the arriving NDT packet can be stored in the queue where it replaces an existing DT packet. Particularly, the DT packet at the rear end of the relay's buffer is pushed out of the queue to make room for the incoming NDT packet, which is inserted behind all of the NDT packets (if any) already present in the buffer. Therefore, if a NDT packet arrives at R, it will be lost only if the buffer is completely filled with NDT packets and the R-D link is in outage resulting in an unsuccessful clearance of one of the NDT packets already present in the buffer. Consequently, the NDT PL can be evaluated as follows:

$$P_{\text{loss}}^{(NDT)} = p_{NDT}p + \alpha_{NDT}\pi_{0,L}q = p_{NDT}p + \alpha_{NDT}\pi_L^{(NDT)}q, \tag{10}$$

where $\pi_{0,L} = \pi_L^{(NDT)}$ since $l_{NDT} = L$ implies that $l_{DT} = 0$ with no uncertainty. Replacing $\pi_L^{(NDT)}$ by its value from (8)-(9) results in an exact closed-form evaluation of (10).

Following from Little' law [25], the APD of the NDT packets can be calculated from:

$$D^{(NDT)} = \frac{\bar{L}_{NDT}}{\eta_{NDT}} + 1, \tag{11}$$

where $\bar{L}_{NDT}$ stands for the average number of NDT packets in the queue while η_{NDT} stands for the output throughput of NDT packets at R. The first term in (11) results from the queuing delay at R since several attempts might be needed to successfully transmit a packet along the R-D link while the addition of a value of 1 results from the deterministic delay of one time slot required for delivering the packet from S to R. Following from (8)-(9) and after straightforward calculations, the average length $\bar{L}_{NDT} = \sum_{l=0}^{L} l\pi_l^{(NDT)}$ can be determined as follows:

$$\bar{L}_{NDT} = \left[\frac{Lr^{L+2} - (L+1)r^{L+1} + r}{r^{L+2} - r^{L+1} - (1-q)r^2 + (1-2q)r + q}\right] \quad ; \quad r = \frac{q\alpha_{NDT}}{(1-q)(1-\alpha_{NDT})}. \tag{12}$$

The throughput η_{NDT} can be determined from the following expression:

$$\eta_{NDT} = \alpha_{NDT}\left(1 - \pi_{0,L}q\right) = \alpha_{NDT}\left(1 - \pi_L^{(NDT)}q\right), \tag{13}$$

where the value of $\pi_L^{(NDT)}$ from (8)-(9) can be replaced in (13). Equation (13) follows since a NDT packet arriving at R (with probability α_{NDT}) can enter the buffer as long as the buffer is not filled with NDT packets or at least one packet exited at the beginning of the time slot.

Three events can trigger the loss of DT packets. (i): The outage of the S-R link implying that a DT packet generated at S (with probability $1-p_{NDT}$) can not reach R. (ii): The arrival of a DT packet at a buffer that is full with the R-D link being in outage. The corresponding probability is $\alpha_{DT}\pi_L^{(tot)}q$ where the specific numbers of NDT and DT packets in the buffer are not important since the arriving packet has a low priority. (iii): The third source of DT packet loss corresponds to the event of dropping a low priority DT packet from a full queue so that a high priority incoming NDT packet can be inserted into the buffer. The corresponding probability is $\alpha_{NDT}(\pi_L^{(tot)} - \pi_L^{(NDT)})q$, where $\pi_L^{(tot)} - \pi_L^{(NDT)}$ is the probability of having a full buffer with at least one DT packet. As a conclusion, DT packets are lost with the following probability:

$$P_{\text{loss}}^{(DT)} = (1 - p_{NDT})p + \alpha_{DT}\pi_L^{(tot)}q + \alpha_{NDT}(\pi_L^{(tot)} - \pi_L^{(NDT)})q. \tag{14}$$

Since $l_{tot} = l_{DT} + l_{NDT}$, then $\bar{L}_{tot} = \bar{L}_{DT} + \bar{L}_{NDT}$, where $\bar{L}_{tot}$ stands for the average number of packets present in the queue (regardless of their type). Following from the similarity of the distributions of l_{NDT} and l_{tot} from (8)-(9), $\bar{L}_{tot}$ can be determined from (12) for the value of r given by $r = \frac{q(1-p)}{(1-q)p}$. Therefore, the average number of DT packets in the queue can be determined from:

$$\bar{L}_{DT} = \left[\frac{Lr_1^{L+2} - (L+1)r_1^{L+1} + r_1}{r_1^{L+2} - r_1^{L+1} - (1-q)r_1^2 + (1-2q)r_1 + q}\right] - \left[\frac{Lr_2^{L+2} - (L+1)r_2^{L+1} + r_2}{r_2^{L+2} - r_2^{L+1} - (1-q)r_2^2 + (1-2q)r_2 + q}\right], \tag{15}$$

where $r_1 = \frac{q(1-p)}{(1-q)p}$ and $r_2 = \frac{q\alpha_{NDT}}{(1-q)(1-\alpha_{NDT})}$.

The output throughput of the DT packets can be determined from:

$$\eta_{DT} = \alpha_{DT} - \alpha_{DT}\pi_L^{(tot)}q - \alpha_{NDT}(\pi_L^{(tot)} - \pi_L^{(NDT)})q, \tag{16}$$

where the last two terms capture the reduction in the effective arrival rate that results from dropping the DT packets due to either: *a)* the relay's buffer being full or *b)* the arrival of a NDT packet at a full relay's buffer having at least one DT packet.

Next, we evaluate the APD for the DT packets, which we denote by $D^{(DT)}$. The preemptive nature of the considered finite capacity queueing system at R renders the analysis of $D^{(DT)}$ more involved as compared to that pertaining to NDT packets. Hence, we provide in Appendix B a detailed derivation of $D^{(DT)}$, whose closed-form expression is found to be given by:

$$D^{(DT)} = 1 + \frac{1}{1 - q\pi_L^{(tot)}} \times \sum_{i=1}^{L} \frac{\left[((q-1)\delta(i-1)+1)\pi_{i-1}^{(tot)} + (1-q)\pi_i^{(tot)}\right]}{a_{i0}} \sum_{j=1}^{L} a_{j0}s_{ij}, \tag{17}$$

where a_{i0} and s_{ij} are defined in (38) and (39) provided in Appendix B. Moreover, $\delta(i-1)$ is the function that is equal to 0 if $i = 1$ and to 1 otherwise.

D. Asymptotic Analysis

In this section, we carry out an asymptotic analysis that holds for $P_M \gg 1$..

Proposition *2:* The steady-state distribution in (7) tends asymptotically to:

$$\begin{cases} (\pi_{1,0}, \pi_{0,1}) \to (1 - p_{NDT}, p_{NDT}), & p > q; \\ (\pi_{L,0}, \pi_{L-1,1}) \to (1 - p_{NDT}, p_{NDT}), & p < q. \end{cases} \quad ; \quad \pi_{l_{DT},l_{NDT}} \to 0 \text{ otherwise.} \tag{18}$$

Proof: The proof is provided in Appendix C. ■

Equation (18) shows that the system performance depends on whether R is closer to D or closer to S. Even though the outage probability of buffer-free relaying systems is minimized for $p = q$ (i.e. $d_{\text{SR}} = d_{\text{RD}}$ for symmetric scintillation along the two hops), other asymmetric relay placements (i.e. $d_{\text{SR}} > d_{\text{RD}}$ and $d_{\text{SR}} < d_{\text{RD}}$) need to be considered for the following reasons. In fact, relays correspond to either dedicated or undedicated nodes. Dedicated relays are deployed by the system engineers with the sole objective of relaying information from one node (S) to another node (D). On the other hand, undedicated relays correspond to the transceivers of neighboring users that are not deployed for the sake of assisting S in its communication with D, but they are independent entities that have their own data to communicate. Since these relays are in the geographical vicinity of S and D, then they can assist the communication between S and D if they have no data to communicate. While the positions of the dedicated relays can be optimized, exploiting neighboring users as relays constitutes a cost effective solution that uses the existing network infrastructure. Even though the placement of dedicated relays can be controlled, yet the relation $d_{\text{SR}} = d_{\text{RD}}$ might be hard to achieve because of the presence of obstacles (that hinder the LOS FSO communications) or because the equidistant point might not be geographically feasible (it occurs in a river or road, for example). Moreover, since adjacent users can be located at arbitrary positions, then $d_{\text{SR}} > d_{\text{RD}}$ or $d_{\text{SR}} < d_{\text{RD}}$ if the user cooperation strategy is to be implemented with the existing infrastructure. It is also important to highlight that the theoretical condition $d_{\text{SR}} = d_{\text{RD}}$ is almost impossible to realize in realistic networks. In fact, these distances are in the order of few kilometers and shifting the relay's location by a few meters while deploying the network will favor lower outages along one of the two hops compared to the other hop.

1) $p > q$ (i.e. $d_{SR} > d_{RD}$): Equation (18) shows that $\pi_L^{(NDT)} \to 0$ and $\pi_L^{(tot)} \to 0$. Replacing these values in (10) and (14) results in:

$$\begin{aligned} &P_{\text{loss}}^{(NDT)} \to p_{NDT} p = p_{NDT} \max\{p, q\}, \\ &P_{\text{loss}}^{(DT)} \to (1 - p_{NDT}) p = (1 - p_{NDT}) \max\{p, q\}. \end{aligned} \quad (19)$$

Similarly, replacing $\pi_L^{(NDT)} \to 0$ and $\pi_L^{(tot)} \to 0$ in (13) and (16) results in $\eta_{NDT} \to \alpha_{NDT} = (1-p)p_{NDT} \to p_{NDT}$ and $\eta_{DT} \to \alpha_{DT} = (1-p)(1-p_{NDT}) \to 1-p_{NDT}$. Finally, from (18), $\pi_0^{(NDT)} \to 1-p_{NDT}$ and $\pi_1^{(NDT)} \to p_{NDT}$, while $\pi_1^{(tot)} \to 1$. This results in $\bar{L}_{NDT} \to p_{NDT}$ and $\bar{L}_{tot} \to 1$ implying that $\bar{L}_{DT} \to 1 - p_{NDT}$ from (15). Consequently, the delays in (11) and (17) tend to the following asymptotic values, clearly highlighting the applicability of Little's law to the calculation of the APD of DT packets in this case:

$$D^{(NDT)} \to 2 \quad ; \quad D^{(DT)} \to 2. \quad (20)$$

2) $p < q$ (i.e. $d_{SR} < d_{RD}$): Equation (18) shows that $\pi_L^{(NDT)} \to 0$ and $\pi_L^{(tot)} \to 1$. Replacing these values in (10) and (14) results in:

$$\begin{aligned} &P_{\text{loss}}^{(NDT)} \to p_{NDT} p = p_{NDT} \min\{p, q\}, \\ &P_{\text{loss}}^{(DT)} \to (1 - p_{NDT}) p + (1 - p) q \to q = \max\{p, q\}, \end{aligned} \quad (21)$$

while the replacement in (13) and (16) shows that $\eta_{NDT} \to \alpha_{NDT} = (1-p)p_{NDT} \to p_{NDT}$, while $\eta_{DT} \to (1-p)(1 - p_{NDT} - q) \to 1 - p_{NDT}$. From (18), $\bar{L}_{NDT} \to p_{NDT}$ and $\bar{L}_{tot} \to L$ implying that $\bar{L}_{DT} \to L - p_{NDT}$ from (15). Therefore, the asymptotic delays can be obtained from (11) and (17) as follows, demonstrating again the applicability of Little's law to the calculation of the APD of DT packets for high SNR regimes:

$$\begin{aligned} &D^{(NDT)} \to 2, \\ &D^{(DT)} \to \frac{(L-1)}{(1-(1-p)p_{NDT})} + 1 + 1 \to \frac{L - p_{NDT}}{1 - p_{NDT}} + 1. \end{aligned} \quad (22)$$

For $d_{\text{SR}} > d_{\text{RD}}$, from (18), the buffer contains only one packet at steady-state where this packet is a NDT (resp. DT) packet with probability p_{NDT} (resp. $1 - p_{NDT}$). Regardless of its type, this packet will eventually leave the buffer with a high probability during the transmission attempt from R at the beginning of the next time slot since R is closer to D. Therefore, both types of packets will be delivered to D with the same asymptotic delay as predicted from (20) since the DT traffic is not penalized by the presence of the NDT packets. Moreover, since the buffer is never full at steady-state ($\pi_L^{(tot)} \to 0$), then the probability of dropping an incoming packet (whether DT or NDT) is negligible implying that the packet loss is dominated by the outage of the S-R link when the packet cannot reach the buffer. This is demonstrated by (19) where both $P_{\text{loss}}^{(NDT)}$ and $P_{\text{loss}}^{(DT)}$ are proportional to p. As a conclusion, when $d_{\text{SR}} > d_{\text{RD}}$, comparable levels of service are guaranteed for both types of traffic since the buffer is not congested in this case. When $d_{\text{SR}} < d_{\text{RD}}$, the arrival rate at R exceeds the departure rate resulting in a buffer that is full all the time ($\pi_L^{(tot)} \to 1$). Moreover, from (18), the buffer is congested with DT packets where $L-1$ and L DT packets are stored in the buffer with probabilities p_{NDT} and $1 - p_{NDT}$, respectively. This substantial queuing of the DT packets increases the asymptotic delay of this type of packets where $D^{(DT)}$ increases with the buffer size as predicted from (22) while the presence of at most one NDT packet at the head of the queue justifies $D^{(NDT)} = 2$ independently from the buffer size. If no packet is capable of exiting the full queue at the beginning of the next time slot (with probability q), then the arrival of a packet at R will incur the drop of a DT packet justifying the fact that $P_{\text{loss}}^{(DT)}$ is proportional to q in (21). On the other hand, if a NDT packet reaches R (with probability $1-p$), then the probability of dropping this packet is almost zero asymptotically since the number of stored NDT packets is much smaller than L implying that $P_{\text{loss}}^{(NDT)}$ is proportional to p as demonstrated in (21). As a conclusion, when $d_{\text{SR}} < d_{\text{RD}}$, the level of QoS differentiation is more pronounced ($P_{\text{loss}}^{(NDT)} < P_{\text{loss}}^{(DT)}$ and $D^{(NDT)} < D^{(DT)}$) following from giving the NDT packets a higher priority in the congested buffer.

3) Conclusions and impact of relay placement: The following conclusions can be drawn from the performed asymptotic analysis:

- Equations (20) and (22) show that the NDT packets can be delivered with the smallest possible delay value of two time slots regardless of the relay position. Therefore, by prioritizing the NDT packets, the proposed scheme fulfills the target of delivering the NDT packets to D with the best possible APD value.
- The scenario $d_{\text{SR}} > d_{\text{RD}}$ favors the reception of the DT packets with the minimum APD value of two at the expense of penalizing the NDT PL that scales as $\max\{p, q\}$.
- The scenario $d_{\text{SR}} < d_{\text{RD}}$ favors the reception of the NDT packets with a smaller PL that scales as $\min\{p, q\}$ at the expense of increasing the DT APD above two.
- The asymptotic analysis reveals that there is no optimal relay placement where different placements will result in different levels of tradeoff between the performance metrics of the NDT and DT traffics. Denote by region-1 and region-2 the sets of points for which $d_{\text{SR}} > d_{\text{RD}}$ and $d_{\text{SR}} < d_{\text{RD}}$, respectively. While the asymptotic APD remains constant over each region, decreasing d_{SR} in region-1 will concurrently reduce the PL's of the NDT and DT traffics while decreasing this distance in region-2 will decrease $P_{\text{loss}}^{(NDT)}$ at the expense of increasing $P_{\text{loss}}^{(DT)}$. As such, for dedicated relays, the vicinity of S constitutes a feasible region for placing R if the upmost priority is to be given to the NDT traffic at the expense of severely penalizing the DT traffic. The vicinity of the midpoint between S and D also constitutes a valid option if the PL of the NDT traffic is to be compromised for the sake of reducing the PL of the DT traffic. Finally, it is not desirable to place R in the vicinity of D since this will negatively impact the performance of both the NDT and DT traffics.

E. Fairness between NDT and DT Traffics

The proposed BA relaying scheme gives full priority to the NDT traffic while completely overlooking the incurred consequences on the DT traffic. In this section, we introduce the idea of packet mutation for the sake of achieving a certain level of fairness between the DT and NDT packets. As with the initially proposed relaying scheme (with no mutation), NDT packets are still prioritized compared to the DT packets when packet mutation is introduced. However, the level of priority of the NDT (resp. DT) packets is decreased (resp. increased) compared to the BA relaying scheme with no mutation. The level of priority can be controlled by a mutation probability β as will be explained later. In practice, mutation should be introduced only when the delay experienced by the NDT packets is found to be smaller than the maximum tolerable delay needed for the proper delivery of these packets. In this case, the parameter β can be adjusted so that the increased value of the APD of the NDT packets is still below the tolerable threshold implying that the delay requirements of the NDT packets are still met. Therefore, the mutation-based solution can be used to improve the average packet delay experienced by the DT packets without violating the average delay requirements pertaining to the NDT packets.

With packet mutation, a DT packet can be mutated into a NDT packet with probability β thus increasing the priority level of the DT packet by treating it as a NDT packet. Following from this mutation, the packets can be classified into three types at the source. (i): The authentic NDT packets, denoted by NDT_a, generated with probability p_{NDT}. (ii): The authentic DT packets, denoted by DT_a, corresponding to the DT packets that are not mutated. These packets are generated with probability $(1 - p_{NDT})(1 - \beta)$. (iii): The mutated DT packets, denoted by DT_m, that are generated with probability $(1-p_{NDT})\beta$. The queue at R will treat these packets as NDT packets; however, their loss and delay must enter in the count of the DT PL and APD.

Denoting the packet losses in (10) and (14) as $P_{\text{loss}}^{(NDT)}(p_{NDT})$ and $P_{\text{loss}}^{(DT)}(p_{NDT})$, respectively:

$$\begin{cases} p_{NDT_a} + p_{DT_m} = P_{\text{loss}}^{(NDT)}(p_{NDT} + (1 - p_{NDT})\beta) \\ p_{DT_a} = P_{\text{loss}}^{(DT)}(p_{NDT} + (1 - p_{NDT})\beta) \end{cases}, \tag{23}$$

where p_{NDT_a}, p_{DT_m} and p_{DT_a} stand for the losses of the packets that are of type NDT_a, DT_m and DT_a, respectively. Given that the authentic NDT traffic accounts for a ratio of $\mathfrak{r} \triangleq \frac{p_{NDT}}{p_{NDT}+(1-p_{NDT})\beta}$ of the total NDT traffic, then the equations in (23) can be readily solved to yield the following PL expressions in the presence of packet mutation:

$$P_{\text{loss,mut}}^{(NDT)} = p_{NDT_a} = \mathfrak{r} P_{\text{loss}}^{(NDT)}(p_{NDT}^{(m)}), \tag{24}$$

$$\begin{aligned} P_{\text{loss,mut}}^{(DT)} &= p_{DT_m} + p_{DT_a} \\ &= (1 - \mathfrak{r}) P_{\text{loss}}^{(NDT)}(p_{NDT}^{(m)}) + P_{\text{loss}}^{(DT)}(p_{NDT}^{(m)}), \end{aligned} \tag{25}$$

where $p_{NDT}^{(m)} \triangleq p_{NDT} + (1-p_{NDT})\beta$ stands for the probability of having a NDT packet when the packet mutation strategy is applied.

On the other hand, denoting the input throughput in (13) as $\eta_{NDT}(p_{NDT}^{(m)})$, then the input throughputs of the NDT_a and DT_m packets are $\mathfrak{r}\eta_{NDT}(p_{NDT}^{(m)})$ and $(1 - \mathfrak{r})\eta_{NDT}(p_{NDT}^{(m)})$, respectively. Similarly, the average queue lengths of these types of packets can be determined from $\mathfrak{r}\bar{L}_{NDT}(p_{NDT}^{(m)})$ and $(1 - \mathfrak{r})\bar{L}_{NDT}(p_{NDT}^{(m)})$, respectively, following from (12). Moreover, denoting the throughput given in (16) as $\eta_{DT}(p_{NDT}^{(m)})$, the throughput of DT_a is $\eta_{DT}(p_{NDT}^{(m)})$. So, $\frac{(1-\mathfrak{r})\eta_{NDT}(p_{NDT}^{(m)})}{(1-\mathfrak{r})\eta_{NDT}(p_{NDT}^{(m)})+\eta_{DT}(p_{NDT}^{(m)})}$ of the total DT packets received at D are DT_m ones that experience an average delay equivalent to NDT packets, while the remaining $\frac{\eta_{DT}(p_{NDT}^{(m)})}{(1-\mathfrak{r})\eta_{NDT}(p_{NDT}^{(m)})+\eta_{DT}(p_{NDT}^{(m)})}$ packets are DT_a ones experiencing a delay equivalent to DT packets.

In light of the above discussion, the APDs with mutation can be determined from:

$$D_{mut}^{(NDT)} = D^{(NDT)}(p_{NDT}^{(m)}) \tag{26}$$

$$= \frac{\mathfrak{r}\bar{L}_{NDT}(p_{NDT}^{(m)}) + (1 - \mathfrak{r})\bar{L}_{NDT}(p_{NDT}^{(m)})}{\mathfrak{r}\eta_{NDT}(p_{NDT}^{(m)}) + (1 - \mathfrak{r})\eta_{NDT}(p_{NDT}^{(m)})} + 1, \tag{27}$$

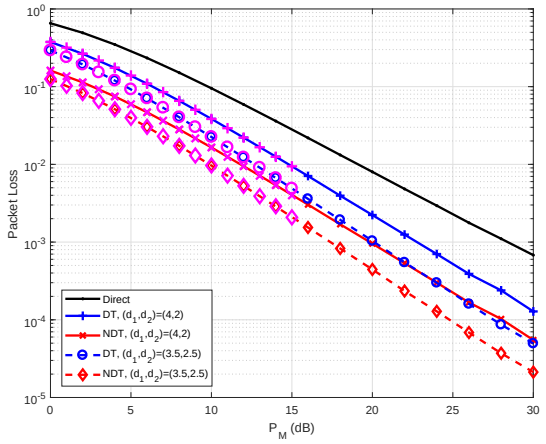

Fig. 3. Packet losses when R is closer to D. Solid lines correspond to the theoretical results while the markers (with no lines) correspond to the numerical results.

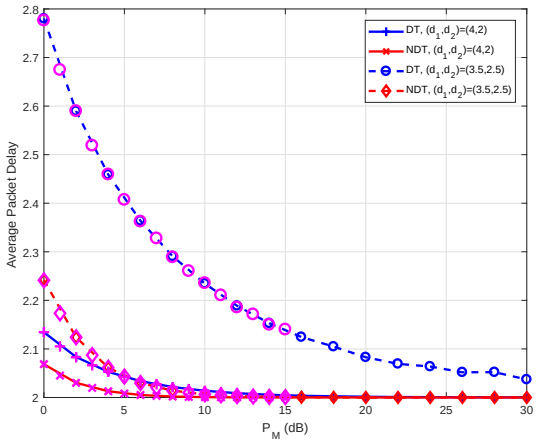

Fig. 4. Average packet delays when R is closer to D. Solid lines correspond to the theoretical results while the markers (with no lines) correspond to the numerical results.

and

$$D_{mut}^{(DT)} = \left(\frac{(1-\mathfrak{r})\eta_{NDT}(p_{NDT}^{(m)})}{(1-\mathfrak{r})\eta_{NDT}(p_{NDT}^{(m)})+\eta_{DT}(p_{NDT}^{(m)})}\right) D^{(NDT)}(p_{NDT}^{(m)}) + \left(\frac{\eta_{DT}(p_{NDT}^{(m)})}{(1-\mathfrak{r})\eta_{NDT}(p_{NDT}^{(m)})+\eta_{DT}(p_{NDT}^{(m)})}\right) D^{(DT)}(p_{NDT}^{(m)}). \quad (28)$$

where the delays $D^{(NDT)}(p_{NDT})$ and $D^{(DT)}(p_{NDT})$ are given in (11) and (17), respectively.

Replacing equations (19) and (21) in (24)-(25) as well as equations (20) and (22) in (27)-(28) shows that the asymptotic PL and APD expressions in (19)-(22) hold in the case of mutation as well. In other words, packet mutation maintains the same performance levels as the no mutation scheme for both DT and NDT packets under the high SNR regime. Nonetheless, the benefits of packet mutation for low SNR values will be illustrated in the next section.

IV. Numerical Results

The refractive index structure constant and the attenuation constant are set to $C_n^2 = 1.7\times10^{-14}$ m$^{-2/3}$ and $\sigma = 0.44$ dB/km, respectively. Unless stated otherwise, we assume that $L = 5$ and $p_{NDT} = 0.3$. The relay is placed along the line joining S with D and its position is determined by the vector $(d_1, d_2) = (d_{\text{SR}}, d_{\text{RD}})$ (all distances will be expressed in km).

Fig. 3 and Fig. 4 show the PL and APD performance, respectively, when R is closer to D where we consider the two scenarios $(d_1, d_2) = (4, 2)$ and $(d_1, d_2) = (3.5, 2.5)$ for a total link distance of 6 km. Results show the close match between the theoretical and numerical results thus highlighting on the validity of the presented performance analysis. In particular, the exact expressions in (10), (11), (14), and (17) are undistinguishable from the numerical results. At high SNRs, the PL and APD curves match the asymptotic values provided in (19) and (20), respectively. In particular, the diversity advantages are the same for the NDT and DT packets where the PL curves are practically parallel to each other at high SNR. Similarly, the APDs tend asymptotically to the optimal value of 2. While the scenario $(d_1, d_2) = (3.5, 2.5)$ results in smaller PLs, the scenario $(d_1, d_2) = (4, 2)$ results in smaller APDs. Results show that in the considered scenarios where the buffer is not congested, the APDs of the NDT and DT packets are comparable especially for $(d_1, d_2) = (4, 2)$.

Fig. 5 and Fig. 6 show the PL and APD performance, respectively, when R is closer to S where we consider the two scenarios $(d_1, d_2) = (2, 4)$ and $(d_1, d_2) = (2.5, 3.5)$ for a total link distance of 6 km. As before, the numerical and theoretical curves overlap for the NDT PL, NDT APD, DT PL, and DT APD. For high values of P_M, the asymptotic APD values are as predicted by (22). Results in Fig. 5 confirm the findings in (21) where the NDT packets profit from a higher diversity advantage compared to the DT packets as evidenced by the slopes of the PL curves. In this scenario, the proposed relaying scheme is clearly privileging the NDT traffic where the performance gap is in the order of 18 dB at a PL of 10^{-3} for $(d_1, d_2) = (2, 4)$. Finally, the QoS differentiation is highly dependent on the relay position. While the NDT traffic for the scenario $(d_1, d_2) = (2, 4)$ experiences the minimum loss and delay, this improvement is realized by compromising the DT traffic that experiences the highest PL and APD among all packet types for all scenarios.

Fig. 7 and Fig. 8 highlight the impact of the buffer size on the PL and APD, respectively, where we compare the performance with $L = 4$ and $L = 8$. For $(d_1, d_2) = (2.5, 2)$ (i.e. $d_{\text{SR}} > d_{\text{RD}}$), results show that the buffer size does not affect the PL and APD of both the DT and NDT packets in coherence with (19) and (20). In fact, (18) shows that in this case the buffer contains only one packet almost always asymptotically at steady-state. As such, adding extra space in the buffer by increasing L from 4 to 8 does not noticeably affect the performance since this extra space is rarely occupied. Results in Fig. 7 and Fig. 8 show that these findings hold for practically all values of P_M. For $(d_1, d_2) = (2, 2.5)$ (i.e.

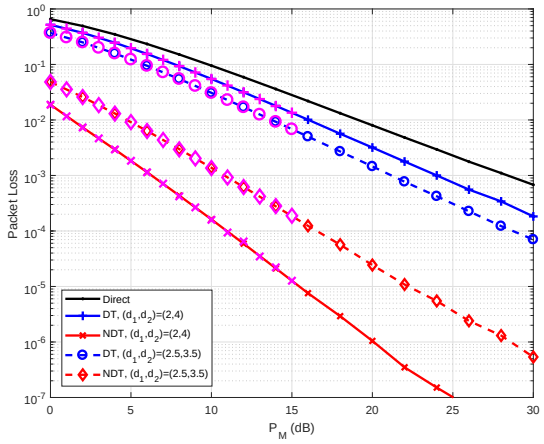

Fig. 5. Packet losses when R is closer to S. Solid lines correspond to the theoretical results while the markers (with no lines) correspond to the numerical results.

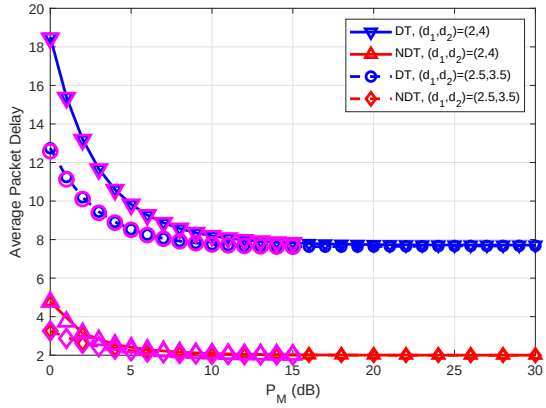

Fig. 6. Average packet delays when R is closer to S. Solid lines correspond to the theoretical results while the markers (with no lines) correspond to the numerical results.

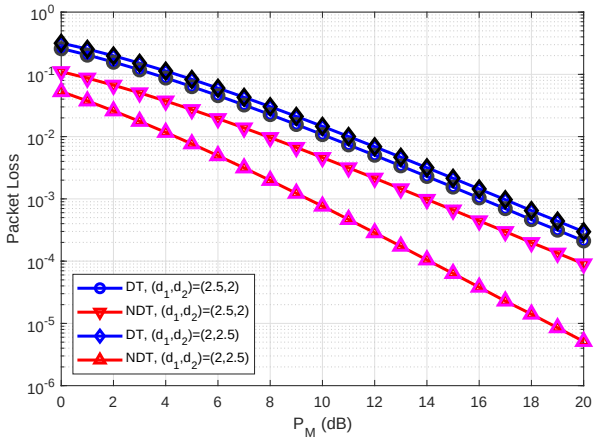

Fig. 7. Impact of the buffer size on the packet loss. Solid lines and markers (with no lines) correspond to the cases $L = 4$ and $L = 8$, respectively.

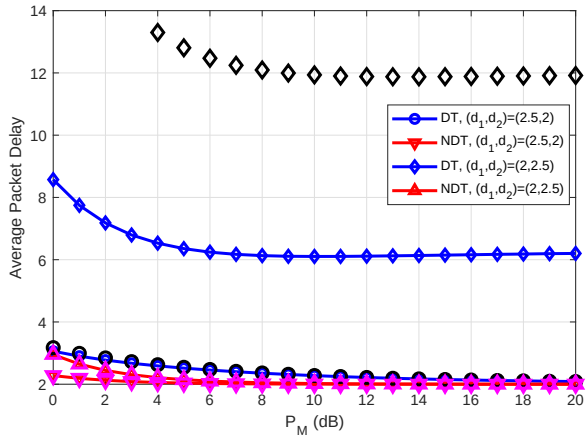

Fig. 8. Impact of the buffer size on the average packet delay. Solid lines and markers (with no lines) correspond to the cases $L = 4$ and $L = 8$, respectively.

$d_{\text{SR}} < d_{\text{RD}}$), results show that the PL is practically not affected by varying L in coherence with (21). In this case, as predicted by (22), the APD of the NDT packets does not vary with L while the APD of the DT packets increases linearly with L. This is demonstrated in Fig. 8 where the asymptotic DT delay increases from 6.28 for $L = 4$ to 12 for $L = 8$, thus, validating (22). The justifications behind these observations are as follows. From (18), there is at most one NDT packet in the buffer when $d_{\text{SR}} < d_{\text{RD}}$ and, hence, all buffer sizes exceeding one will impact the NDT traffic in the same way. On the other hand, the buffer is full and congested with DT packets according to (18). As such, an arriving low priority DT packet will be dropped and an arriving high priority NDT packet will trigger the drop of a DT packet. Such scenarios will arise independently from the size of the buffer that will always be filled at steady-state (regardless of its size). Consequently, the PL curves of the DT packets almost overlap for $L = 4$ and $L = 8$ since the drop of four additional DT packets in the former case will have an unremarkable impact on the continuous packet flow. On the other hand, the longer the queue, the more time the DT packets (whose number is either $L-1$ or L from (18)) will remain in the queue, thus, negatively impacting the DT delay.

Fig. 9 and Fig. 10 show the impact of packet mutation for $(d_1, d_2) = (2, 3)$ and $\beta = 0, 0.4, 0.8$. It is worth highlighting that for $\beta = 0$, the relaying scheme with mutation simplifies to the initially proposed BA priority scheme (with no mutation) in Section III-A where the full priority is given to the NDT packets while completely ignoring the corresponding consequences on the DT traffic. In this context, it can be observed that setting $\beta = 0$ in (24)-(28) results in $P_{\text{loss,mut}}^{(NDT)} = P_{\text{loss}}^{(NDT)}$, $P_{\text{loss,mut}}^{(DT)} = P_{\text{loss}}^{(DT)}$, $D_{mut}^{(NDT)} = D^{(NDT)}$ and $D_{mut}^{(DT)} = D^{(DT)}$ since $\mathfrak{r} = 1$ for $\beta = 0$. As highlighted in Section III-E, the asymptotic DT PL and APD values are not improved relative to the no-mutation strategy. However, huge gains are observed for DT packets in terms of delay performance for low SNR values. This is achieved at the cost of an increase in the delay experienced by NDT packets. For $P_M = 7$ dB, results in Fig. 10 show that the initially proposed relaying scheme with no packet mutation guarantees an APD of 2.1 for the

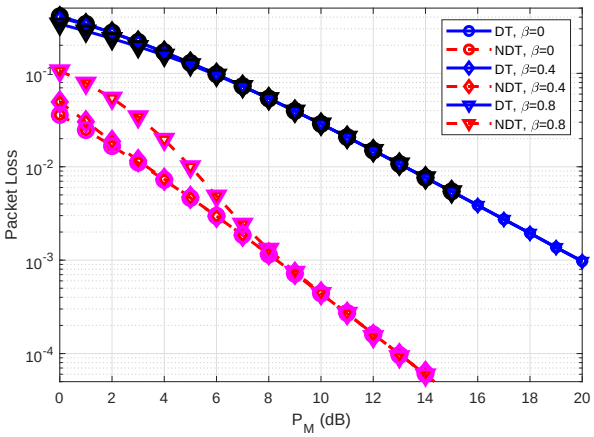

Fig. 9. Impact of mutation on the packet loss for $(d_1, d_2) = (2, 3)$. Solid lines correspond to the theoretical results while the markers (with no lines) correspond to the numerical results.

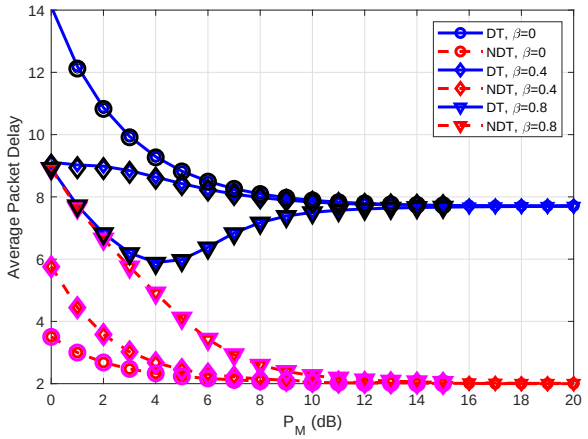

Fig. 10. Impact of mutation on the APD for $(d_1, d_2) = (2, 3)$. Solid lines correspond to the theoretical results while the markers (with no lines) correspond to the numerical results.

NDT packets. Assuming, for example, that the NDT traffic can tolerate a delay up to 3, then the packet mutation strategy can be applied with $\beta = 0.8$. From Fig. 10, this approach increases the APD of the NDT packets from 2.1 to 2.9 while decreasing the APD of the DT packets from 8.3 to 6.8. In other words, the reduction in the delay of the DT packets was achieved while maintaining the delay of the NDT packets below the tolerated limit of 3. From Fig. 9, the PL of the NDT packets was slightly affected by the introduced packet mutation.

V. Conclusion

In this paper, we proposed a novel relaying/buffering scheme suitable for full-duplex FSO communications with NDT and DT traffics. The results obtained from both Markov chain analysis and simulations demonstrated the ability of the proposed priority queueing discipline to achieve lower delays for NDT packets as compared to DT packets. This allows for a better enforcement of QoS differentiation in the context of BA FSO communication systems. Class of service mutation was also introduced as a means for strategically mitigating the effect of NDT packets on DT ones. Finally, an asymptotic analysis highlighted on the impact of the relay placement on the achievable diversity gains and APD values. This analysis also showed the capability of the proposed priority-buffering scheme in guaranteeing small delay values for the delay-sensitive NDT traffic even in the presence of class of service mutation.

Appendix A

Consider a buffer of size L whose dynamics are similar to the dynamics of the NDT packets described in Section III-A. Denote by λ the probability of a packet arriving at the queue and by $\mu = 1-q$ the probability of a packet leaving the queue. Denoting by l the actual number of packets in the queue, the following cases arise.

Case 1: $l = 0$. For an empty queue, no packet can be transmitted to D implying that:

$$t_{0,0} = 1 - \lambda \;\; ; \;\; t_{0,1} = \lambda. \tag{29}$$

Case 2: $l = L$. For a full queue, the following transitions are possible:

$$t_{L,L} = (1-\mu) + \mu\lambda \;\; ; \;\; t_{L,L-1} = \mu(1-\lambda) \triangleq c_{-1}, \tag{30}$$

where, for the calculation of $t_{L,L}$, either (i): no packet can be transmitted (with probability $1-\mu$) implying that no packet can be received since the buffer remains full or (ii): a packet can be transmitted (with probability μ) implying that an incoming packet (with probability λ) will make the buffer full again. With probability $\mu(1-\lambda)$ one packet is transmitted while no packet is received thus reducing the number of packets by one.

Case 3: For $l \neq 0$ and $l \neq L$, the number of packets in the buffer can remain the same, increase by one or decrease by one according to the following probabilities:

$$t_{l,l} = \mu\lambda + (1-\mu)(1-\lambda) \;\; ; \;\; t_{l,l-1} = \mu(1-\lambda) = c_{-1} \;\; ; \;\; t_{l,l+1} = (1-\mu)\lambda \triangleq c_1. \tag{31}$$

We denote by $\mathfrak{p}_l$ the probability of having l packets in the queue at steady-state. The balance equation at the state $l = 0$ is given by:

$$\lambda\mathfrak{p}_0 = c_{-1}\mathfrak{p}_1 \;\Rightarrow\; \frac{\mathfrak{p}_1}{\mathfrak{p}_0} = \frac{\lambda}{\mu(1-\lambda)} = \frac{r}{1-\mu}, \tag{32}$$

where $r \triangleq \frac{(1-\mu)\lambda}{\mu(1-\lambda)} = \frac{q\lambda}{(1-q)(1-\lambda)}$.

At $l = 1$, $(c_{-1} + c_1)\mathfrak{p}_1 = \lambda\mathfrak{p}_0 + c_{-1}\mathfrak{p}_2$ which, from (32), results in $\frac{\mathfrak{p}_2}{\mathfrak{p}_1} = \frac{c_1}{c_{-1}} = r$. Similarly, the recursive application of the balance equations for $l = 2, \ldots, L-1$ results in $(c_{-1} + c_1)\mathfrak{p}_l = c_{-1}\mathfrak{p}_{l+1} + c_1\mathfrak{p}_{l-1}$ implying that:

$$\frac{\mathfrak{p}_2}{\mathfrak{p}_1} = \frac{\mathfrak{p}_3}{\mathfrak{p}_2} = \cdots = \frac{\mathfrak{p}_L}{\mathfrak{p}_{L-1}} = r. \tag{33}$$

Combining (32) and (33) results in $\mathfrak{p}_l = \frac{r^l}{1-\mu}\mathfrak{p}_0$ for $l = 1, \ldots, L$. Following from this relation and the fact that

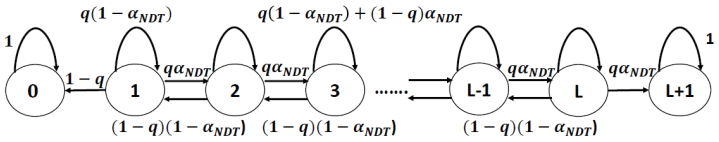

Fig. 11. The random walk experienced by a DT packet at R.

$\sum_{l=0}^{L} \mathfrak{p}_l = 1$ while using the geometric series sum formula results in the solution provided in (8). Finally, observing that a NDT packet (resp. packet of any type) arrives at the queue with probability α_{NDT} (resp. $1-p$) results in (9).

APPENDIX B

In this appendix, we calculate $D^{(DT)}$, the mean waiting time of a DT packet given that it got accepted into the relay's queue and it was successfully transmitted to D. Note that the latter event is referred to in what follows as a packet service event. $D^{(DT)}$ has two components, namely one component capturing the deterministic delay of one time slot at S, along with another one representing the delay at R, which we denote by $D_R^{(DT)}$. The approach used to obtain $D_R^{(DT)}$, and hence $D^{(DT)}$, is built upon the one-dimensional random walk depicted in Fig. 11, and which describes the evolution of the position of a DT packet within the relay's queue over time.

The different states that make up the random walk can be interpreted as follows. We tag an arriving DT packet when it enters the relay's queue. The state of the tagged DT packet is characterized by its position during the course of its residency in the relay's queue. Note that in the state transition diagram portrayed in Fig. 11, two special states stand out, namely the states of 0 and $L+1$. By convention, a state of 0 for the tagged DT packet means that the packet got served (i.e., is sent to D). Moreover, the random walk reaches the absorbing state $L+1$, when the tagged DT packet is pushed out of the relay's queue by the subsequently arriving NDT packets, leading thus to a DT packet loss event. If upon arrival at R, the tagged DT packet finds i NDT packets and j DT packets in the queue, then the random walk starts at state $i+j+1$.

Clearly, the position of the tagged DT packet is only affected by the arrival of NDT packets. Given that at most one NDT packet can be received per time slot, transitions between adjacent states of the considered random walk are governed by the following three cases. Case 1: The tagged DT packet moves forward in the queue, if no NDT packet is received and the packet at the front of the relay's queue gets served. Case 2: The tagged DT packet moves backwards in the queue, if a NDT packet is received and the front packet is not served. Case 3: The tagged DT packet maintains the same position if (i): neither an NDT packet arrives at R nor a departure to D occurs or (ii): an NDT packet arrives at R and the front packet gets served.

Let $P_{i\to j}$ be the single step transition probability from position i to position j of the considered random walk. The transition probabilities corresponding to the three above listed cases can be found as follows, for $2 \leq i \leq L$. Case 1: $P_{i\to i-1} = (1-\alpha_{NDT})(1-q)$. Case 2: $P_{i\to i+1} = q\alpha_{NDT}$. Case 3: $P_{i\to i} = q(1-\alpha_{NDT})+(1-q)\alpha_{NDT}$. Note that for the special case of $i=1$, we have $P_{1\to 2} = q\alpha_{NDT}$, $P_{1\to 0} = 1-q$ and $P_{1\to 1} = 1 - P_{1\to 0} - P_{1\to 2}$.

The previously obtained transition probabilities can be used to construct the state transition matrix of the considered random walk and that we denote by $\mathbf{P}_{(L+2)\times(L+2)}$. This matrix is of particular interest as it will be instrumental in finding the mean waiting time of a tagged DT packet that enters the relay's queue and gets served; that is, $D_R^{(DT)}$. This objective can be achieved by deriving the mean time to absorption into the state 0 of the random walk. The latter metric represents the average number of time slots required by the tagged DT packet to reach the absorbing state 0 and go thus into service. Through a simple rearrangement of the states in $\mathbf{P}$ to have all L transient states of the random walk (namely, the ones corresponding to states 1 through L) precede the two absorbing states 0 and $L+1$, $\mathbf{P}$ would take the following form:

$$\mathbf{P}_{(L+2)\times(L+2)} = \left(\begin{array}{c|c} \mathbf{Q}_{L\times L} & \mathbf{R}_{L\times 2} \\ \hline \mathbf{0}_{2\times L} & \mathbf{I}_{2\times 2} \end{array}\right), \tag{34}$$

where $\mathbf{I}_{2\times 2}$ is a 2×2 identity matrix, $\mathbf{0}_{2\times L}$ is a $2\times L$ matrix whose components are all equal to 0, $\mathbf{Q}_{L\times L}$ is a $L\times L$ matrix capturing the transitions among the transient states, and $\mathbf{R}_{L\times 2}$ is a matrix reflecting the transitions between transient and absorbing states of the random walk. The fact that the last two rows/columns of the $\mathbf{P}$ matrix are reserved for the two absorbing states 0 and $L+1$ justifies the existence of the $\mathbf{I}_{2\times 2}$ and $\mathbf{0}_{2\times L}$ matrices in the lower part of $\mathbf{P}$.

It is well known that, in the context of an absorbing random walk having a finite number of states, the matrix $\mathbf{I}_{L\times L}-\mathbf{Q}_{L\times L}$ is invertible [26]. Given the transition probabilities shown in Fig. 11, $\mathbf{I}_{L\times L} - \mathbf{Q}_{L\times L}$ is a tridiagonal matrix with the following elements: $(\mathbf{I}-\mathbf{Q})_{i,i+1} = -q\alpha_{NDT}$, $(\mathbf{I}-\mathbf{Q})_{i,i-1} = -(1-q)(1-\alpha_{NDT})$, $(\mathbf{I}-\mathbf{Q})_{1,1} = 1-q(1-\alpha_{NDT})$, $(\mathbf{I}-\mathbf{Q})_{i,i} = 1-q(1-\alpha_{NDT})-(1-q)\alpha_{NDT}$ for $i\neq 1$, and $(\mathbf{I}-\mathbf{Q})_{i,j} = 0$ for $|i-j|>1$.

$\mathbf{S}_{L\times L} = (\mathbf{I}_{L\times L}-\mathbf{Q}_{L\times L})^{-1}$ is referred to as the fundamental matrix and its (i,j)-th element, s_{ij}, represents the expected number of time slots the random walk is in transient state j starting initially from the transient state i. This renders $\mathbf{S}_{L\times L}$ central in computing one of the fundamental performance metrics in our analysis, in particular, the mean time to absorption into state 0. Owing to the fact that $\mathbf{I}_{L\times L}-\mathbf{Q}_{L\times L}$ is tridiagonal and as stipulated in [27] where a formula for the inverse of a general tridiagonal matrix is provided, s_{ij} can be determined as follows:

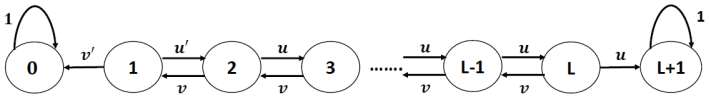

Fig. 12. An embedded Markov chain model of the random walk experienced by a DT packet at R.

$$s_{ij} = \begin{cases} \frac{(q\alpha_{NDT})^{j-i}\theta_{i-1}\phi_{j+1}}{\theta_L}, & \text{if } i < j \\ \frac{\theta_{i-1}\phi_{j+1}}{\theta_L}, & \text{if } i = j \\ \frac{((1-q)(1-\alpha_{NDT}))^{i-j}\theta_{j-1}\phi_{i+1}}{\theta_L} & \text{if } i > j \end{cases}, \quad (35)$$

where θ_i satisfies the recurrence relation: $\theta_i = (1 - q(1 - \alpha_{NDT}) - (1-q)\alpha_{NDT})\theta_{i-1} - q\alpha_{NDT}(1-q)(1-\alpha_{NDT})\theta_{i-2}$ for $2 \leq i \leq L$, with the initial conditions $\theta_0 = 1$ and $\theta_1 = 1 - q(1-\alpha_{NDT})$. In addition, ϕ_i follows the recurrence relation: $\phi_i = (1 - q(1 - \alpha_{NDT}) - (1 - q)\alpha_{NDT})\phi_{i+1} - q\alpha_{NDT}(1-q)(1-\alpha_{NDT})\phi_{i+2}$ for $1 \leq i \leq L-1$, with the initial conditions $\phi_{L+1} = 1$ and $\phi_L = 1 - q(1-\alpha_{NDT}) - (1-q)\alpha_{NDT}$. Developing a closed form expression for θ_i and ϕ_i based on their associated recurrence relations is out of reach. However, it is still possible to find θ_i and ϕ_i, and hence s_{ij}, by exploiting the following observation. In fact, according to the guidelines given in [26], the so-called absorption probability matrix $\mathbf{A}_{L\times 2}$, whose elements represent the probability of ending up in one of the two absorbing states, can be obtained as follows:

$$\mathbf{A}_{L\times 2} = \mathbf{S}_{L\times L} \times \mathbf{R}_{L\times 2}. \quad (36)$$

Note that the (i,j)-th element of matrix $\mathbf{A}$, that we denote by a_{ij}, is the probability of getting absorbed into absorbing state $j \in \{0, L+1\}$ coming from transient state i. As such, a_{i0} for $i = 1, \ldots, L$, gives the probability that our tagged DT packet goes into service (i.e., reaches absorbing state 0 of the random walk), starting from transient state i. Similarly, $a_{i,L+1} = 1 - a_{i0}$ for $i = 1, \ldots, L$ represents the probability that the DT packet gets pushed out of the relay's queue. Note that both a_{i0} and s_{ij} are needed to compute the mean waiting time at R of a DT packet that gets served. Given that our aim is to find a closed form expression for the latter quantity, it is necessary to find one for a_{i0} and s_{ij} as well. In the sequel, we derive first an expression for a_{i0}, which we then use to come up with a closed form expression for s_{ij}, based on Eq. (36).

To evaluate a_{i0}, an embedded Markov chain model of the random walk given in Fig. 11 is considered. The embedding points of time are those at which an actual transition out of a state occurs. This enables us to remove the self-loops from the original Markov chain model of the random walk, simplifying thus the analysis. When constructing the embedded Markov chain, the transition probability $P_{i\to j}$ is replaced by $\frac{P_{i\to j}}{1-P_{i\to i}}$ [26]. Letting, for $2 \leq i \leq L$, $u = \frac{P_{i\to i+1}}{1-P_{i\to i}} = \frac{q\alpha_{NDT}}{q\alpha_{NDT}+(1-\alpha_{NDT})(1-q)}$, $v = 1-u$, $u' = \frac{P_{1\to 2}}{1-P_{1\to 1}} = \frac{q\alpha_{NDT}}{q\alpha_{NDT}+1-q}$, and $v' = 1-u'$, we get the embedded Markov chain model of the random walk that is given in Fig. 12.

This results in a special type of random walk that commonly arises in the context of the classical gambler's ruin problem [28]. In the context of our considered problem, a tagged DT packet wins if it goes into service, that is, it reaches absorbing state 0 of the random walk. Based on this observation, a DT packet starting from position i in the relay's queue wins the game with a probability equal to a_{i0}. Given the structure of the embedded Markov chain depicted in Fig. 12, it is clear that a_{i0} satisfies the following recurrence relation:

$$a_{i0} = \begin{cases} u'a_{20} + v', & \text{if } i = 1 \\ ua_{i+1,0} + va_{i-1,0} & \text{if } 2 \leq i \leq L \end{cases}, \quad (37)$$

with $a_{i0} = 1$ for $i = 0$ and $a_{i0} = 0$ for $i = L+1$. Solving (37), using an adapted version of the classical solutions delineated in [28], yields for $1 \leq i \leq L+1$:

$$a_{i0} = \frac{(1-q)^i(1-\alpha_{NDT})^{i-1}}{q^L\alpha_{NDT}^{L+1} - (1-q)^{L+1}(1-\alpha_{NDT})^L} \times \left(q^{L-i+1}\alpha_{NDT}^{L-i+1} - (1-q)^{L-i+1}(1-\alpha_{NDT})^{L-i+1}\right). \quad (38)$$

Having obtained a_{i0} (and $a_{i,L+1} = 1 - a_{i0}$) and following from the relation given in Eq. (36) while capitalizing on the fact that the $\mathbf{R}$ matrix has only two non-zero elements, namely $\mathbf{R}_{1,0} = 1 - q$ and $\mathbf{R}_{L,L+1} = q\alpha_{NDT}$, we get the following expression for s_{ij}:

$$s_{ij} = \begin{cases} \frac{(q\alpha_{NDT})^{j-1}a_{i,L+1}a_{j,0}}{a_{1,L+1}(1-q)^j(1-\alpha_{NDT})^{j-1}}, & \text{if } i \leq j \\ \frac{(q\alpha_{NDT})^{j-1}a_{j,L+1}a_{i,0}}{a_{1,L+1}(1-q)^j(1-\alpha_{NDT})^{j-1}} & \text{if } i > j \end{cases}, \quad (39)$$

Armed with Eqs. (38) and (39), we proceed to calculating $D_R^{(DT)}$ next. Let N_{ij} be a random variable representing the number of visits to state j beginning from state i. As stated previously, given that s_{ij} is the expected number of visits to state j starting from transient state i, it follows that $E[N_{ij}] = s_{ij}$. However, our objective is to find the expected value of N_{ij} given that the random walk reaches absorbing state 0. So, let $G(i)$ be the event that the random walk gets absorbed into state 0 coming from transient state i. Recall that $\Pr[G(i)] = a_{i0}$. Therefore, the quantity of interest can be calculated as follows:

$$E[N_{ij}|G(i)] = \sum_{k=0}^{\infty} k\Pr[N_{ij} = k|G(i)] = \sum_{k=0}^{\infty} k\frac{\Pr[N_{ij} = k, G(i)]}{a_{i0}}. \quad (40)$$

Let f_{ij} be the probability of ever visiting state j coming from state i. As discussed in [26], f_{ij} can be expressed as a function of s_{ij} as follows:

$$f_{ii} = 1 - \frac{1}{s_{ii}} \quad ; \quad f_{ij} = \frac{s_{ij}}{s_{jj}} \quad \text{for} \quad i \neq j. \quad (41)$$

f_{ij} can be used to find $\Pr[N_{ij} = k, G(i)]$ as follows. In point of fact, the event $(N_{ij} = k, G(i))$ denotes the event of visiting k times state j beginning from transient state i, before

absorption into state 0. This reduces to moving from state i to state j, then moving from state j back to state j $(k-1)$ times, before going into absorbing state 0 without visiting state j again. This translates into the following relation:

$$\Pr\left[N_{ij}=k, G(i)\right] = f_{ij} \times f_{jj}^{k-1} \times (1-f_{jj}) \times a_{j0}. \quad (42)$$

By substituting (42) into (40), we get after some straightforward manipulations:

$$E\left[N_{ii}|G(i)\right] = \frac{1}{1-f_{ii}} = s_{ii}, \quad (43)$$

$$E\left[N_{ij}|G(i)\right] = \frac{f_{ij}a_{j0}}{a_{i0}(1-f_{jj})} = \frac{a_{j0}s_{ij}}{a_{i0}} \quad \text{for} \quad i \neq j. \quad (44)$$

Based on the quantities given above, the expected number of steps performed starting from state i before absorption is given by: $E\left[N_i|G(i)\right] = \sum_{j=1}^{L} E\left[N_{ij}|G(i)\right]$. Therefore, it becomes possible to find the mean waiting time of a DT packet at R given that it entered the queue (probability of that event is $1-q\pi_L^{(tot)}$) and got served as follows:

$$D_R^{(DT)} = \frac{1}{1-q\pi_L^{(tot)}} \sum_{i=1}^{L} \Big[((q-1)\delta(i-1)+1)\pi_{i-1}^{(tot)} + (1-q)\pi_i^{(tot)}\Big] E\left[N_i|G(i)\right], \quad (45)$$

where the denominator represents the probability that a DT packet enters the relay's queue, $\delta(i-1)$ is the function that is equal to 0 if $i=1$ and to 1 otherwise, and the $\Big[((q-1)\delta(i-1)+1)\pi_{i-1}^{(tot)} + (1-q)\pi_i^{(tot)}\Big]$ quantity in the numerator is the probability that an admitted DT packet starts its random walk from position i in the relay's queue. By adding the deterministic delay of one time slot to (45), we get the expression for $D^{(DT)}$ in (17).

Appendix C

Consider first the case $p > q$. Ignoring all terms that are multiplied by q while approximating $1-q$ with 1, we focus on the states $\{(0,0),(0,1),(1,0)\}$. From case 1 in Section III-B, the nonzero transition probabilities from the state $(0,0)$ are: $t_{(0,0),(0,0)} \to 1-\alpha_{NDT}-\alpha_{DT}$, $t_{(0,0),(0,1)} \to \alpha_{NDT}$ and $t_{(0,0),(1,0)} \to \alpha_{DT}$. From case 4 in Section III-B, $t_{(0,1),(0,0)} \to 1-\alpha_{NDT}-\alpha_{DT}$, $t_{(0,1),(0,1)} \to \alpha_{NDT}$ and $t_{(0,1),(1,0)} \to \alpha_{DT}$ while all other transition probabilities from the state $(0,1)$ tend to zero. Finally, from case 5 in Section III-B, the possible transitions are $t_{(1,0),(0,0)} \to 1-\alpha_{NDT}-\alpha_{DT}$, $t_{(1,0),(0,1)} \to \alpha_{NDT}$ and $t_{(1,0),(1,0)} \to \alpha_{DT}$.

The above analysis implies that the states $\{(0,0),(0,1),(1,0)\}$ form a closed subset where, asymptotically, no transitions are possible from any state in this set to any state outside the set [26, Sec. 9.5]. Therefore, the probability to be in this set will tend to one where, after a certain number of transitions, the Markov chain will move to this set without the capability of leaving it since the transition probabilities out of this set all tend to zero. Therefore, $\pi_{0,0}+\pi_{0,1}+\pi_{1,0} \to 1$. Solving this equation along with the balance equation results in: $\pi_{0,0} \to 1-\alpha_{NDT}-\alpha_{DT} = p \to 0$, $\pi_{0,1} \to \alpha_{NDT} = (1-p)p_{NDT} \to p_{NDT}$ and $\pi_{1,0} \to \alpha_{DT} = (1-p)(1-p_{NDT}) \to 1-p_{NDT}$.

Similarly, for the case $p < q$, we will prove that the states $\{(L,0),(L-1,1)\}$ form a closed subset. From case 3 in Section III-B, $t_{(L,0),(L,0)} = 1-p_{NDT}$ and $t_{(L,0),(L-1,1)} = p_{NDT}$. Similarly, from case 6 in Section III-B, $t_{(L-1,1),(L,0)} = 1-p_{NDT}$ and $t_{(L-1,1),(L-1,1)} = p_{NDT}$. The above relations imply that $\pi_{L,0} = 1-p_{NDT}$ and $\pi_{L-1,1} = p_{NDT}$ completing the proof.

References

[1] J. Kurose and K. Ross, "Computer networking: a top down approach," *Pearson Education*, 2017.

[2] V. Jungnickel et al, "The role of small cells, coordinated multipoint, and massive MIMO in 5G," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 44 – 51, 2014.

[3] M. Safari and M. Uysal, "Relay-assisted free-space optical communication," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 5441 – 5449, Dec. 2008.

[4] M. Karimi and M. Nasiri-Kenari, "Free-space optical communications via optical amplify-and-forward relaying," *J. Lightwave Technol.*, vol. 29, no. 2, pp. 242–2248, Jan. 2011.

[5] C. Abou-Rjeily, "Achievable diversity orders of decode-and-forward cooperative protocols over gamma-gamma fading FSO links," *IEEE Trans. Commun.*, vol. 61, no. 9, pp. 3919–3930, Sep. 2013.

[6] L. Yang, X. Gao, and M.-S. Alouini, "Performance analysis of relay-assisted all-optical FSO networks over strong atmospheric turbulence channels with pointing errors," *J. Lightwave Technol.*, vol. 32, no. 23, pp. 4613–4620, Dec. 2014.

[7] B. Zhu, J. Cheng, M. Alouini, and L. Wu, "Relay placement for FSO multi-hop DF systems with link obstacles and infeasible regions," *IEEE Trans. Wireless Commun.*, vol. 14, no. 9, pp. 5240 – 5250, Sep. 2015.

[8] N. D. Chatzidiamantis, D. S. Michalopoulos, E. E. Kriezis, G. K. Karagiannidis, and R. Schober, "Relay selection protocols for relay-assisted free-space optical systems," *IEEE J. Opt. Commun. Netw.*, vol. 5, no. 1, pp. 4790 –4807, January 2013.

[9] C. Abou-Rjeily, "Performance analysis of selective relaying in cooperative free-space optical systems," *J. Lightwave Technol.*, vol. 31, no. 18, pp. 2965–2973, September 2013.

[10] A. Bletsas, A. Khisti, D. P. Reed, and A. Lippman, "A simple cooperative diversity method based on network path selection," *IEEE J. Select. Areas Commun.*, vol. 24, no. 3, pp. 659–672, Mar. 2006.

[11] C. Abou-Rjeily, "Performance analysis of FSO communications with diversity methods: Add more relays or more apertures?" *IEEE J. Select. Areas Commun.*, vol. 33, no. 9, pp. 1890 – 1902, Sep. 2015.

[12] I. Krikidis, T. Charalambous, and J. S. Thompson, "Buffer-aided relay selection for cooperative diversity systems without delay constraints," *IEEE Trans. Wireless Commun.*, vol. 11, no. 5, pp. 1957–1967, May 2012.

[13] Z. Tian, G. Chen, Y. Gong, Z. Chen, and J. A. Chambers, "Buffer-aided max-link relay selection in amplify-and-forward cooperative networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 2, pp. 553–565, Feb. 2015.

[14] Z. Tian, Y. Gong, G. Chen, and J. Chambers, "Buffer-aided relay selection with reduced packet delay in cooperative networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2567–2575, Mar. 2017.

[15] B. Manoj, R. K. Mallik, and M. R. Bhatnagar, "Performance analysis of buffer-aided priority-based max-link relay selection in DF cooperative networks," *IEEE Trans. Commun.*, vol. PP, no. 99, pp. 1–1, Oct. 2018.

[16] A. A. M. Siddig and M. F. M. Salleh, "Buffer-aided relay selection for cooperative relay networks with certain information rates and delay bounds," *IEEE Trans. Veh. Technol.*, vol. 66, no. 11, pp. 10499–10514, Nov. 2017.

[17] V. Jamali, D. S. Michalopoulos, M. Uysal, and R. Schober, "Link allocation for multiuser systems with hybrid RF/FSO backhaul: Delay-limited and delay-tolerant designs," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3281–3295, May 2016.

[18] Y. F. El-Eryani, A. M. Salhab, S. A. Zummo, and M. S. Alouini, "Protocol Design and Performance Analysis of Multiuser Mixed RF and Hybrid FSO/RF Relaying with Buffers," *IEEE/OSA Journal of Optical Commun. and Networking*, vol. 10, no. 4, pp. 309–321, April 2018.

[19] M. Najafi, V. Jamali, and R. Schober, "Optimal relay selection for the parallel hybrid RF/FSO relay channel: Non-bufferaided and buffer-aided designs," *IEEE Trans. Wireless Commun.*, vol. 7, no. 65, pp. 2794–2810, April 2017.

[20] M.Z. Hassan, M.J. Hossain, J. Cheng and V.C.M. Leung, "Hybrid RF/FSO backhaul network with statistical-QoS aware buffer-aided relaying," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1464–1483, Mar. 2020.
[21] C. Abou-Rjeily and W. Fawaz, "Buffer-aided relaying protocols for cooperative FSO communications," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 8205–8219, Dec. 2017.
[22] N. Nasser, L. Karim and T. Taleb, "Dynamic multilevel priority packet scheduling scheme for wireless sensor network," *IEEE Trans. Wireless Commun.*, vol. 12, no. 4, pp. 1448–1459, Feb. 2013.
[23] M. Emara, H. El Sawy and G. Bauch, "Prioritized multi-stream traffic in uplink IoT networks: Spatially interacting vacation queues," *IEEE Internet of Things Journal*, vol. 8, no. 3, pp. 1477–1491, Jul. 2020.
[24] C.C Sobin, "A survey of routing and data dissemination in delay tolerant networks," *Elsevier Journal of Network and Computer Applications*, vol. 67, pp. 128–146, May 2016.
[25] J. D. C. Little and S. C. Graves, "Little's law," in *International Series in Operations Research & Management Science*, New York, NY, USA: Springer-Verlag, vol. 115, pp. 81–100, 2008.
[26] W. J. Stewart, *Probability, Markov chains, queues, and simulation: the mathematical basis of performance modeling*. Princeton University Press, July 2009.
[27] R.A. Usmani, "Inversion of a Tridiagonal Jacobi Matrix," *Linear Algebra and its Applications*, vol. 212, no. 213, pp. 413–414, Jun. 1994.
[28] S. Song and J. Song, "A Note on the History of the Gambler's Ruin Problem," *Communications for Statistical Applications and Methods*, vol. 20, no. 2, pp. 157–168, Mar. 2013.