entropy

Article

Broadcast Approach to Uplink NOMA: Queuing Delay Analysis

Maha Zohdy [1,†,‡], Ali Tajer [1,*,†] and Shlomo Shamai (Shitz) [2,†]

[1] Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180, USA
[2] Faculty of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 3200003, Israel
[*] Correspondence: tajer@ecse.rpi.edu; Tel.: +1-518-276-8237
[†] The authors contributed equally to this work.
[‡] This author was with Rensselaer Polytechnic Institute when this work was completed. Currently, she is with MathWorks Inc., Natick, MA 01760, USA.

Abstract: Emerging wireless technologies are envisioned to support a variety of applications that require simultaneously maintaining low latency and high reliability. Non-orthogonal multiple access techniques constitute one candidate for grant-free transmission alleviating the signaling requirements for uplink transmissions. In open-loop transmissions over fading channels, in which the transmitters do not have access to the channel state information, the existing approaches are prone to facing frequent outage events. Such outage events lead to repeated re-transmissions of the duplicate information packets, penalizing the latency. This paper proposes a multi-access broadcast approach in which each user splits its information stream into several information layers, each adapted to one possible channel state. This approach facilitates preventing outage events and improves the overall transmission latency. Based on the proposed approach, the average queuing delay of each user is analyzed for different arrival processes at each transmitter. First, for deterministic arrivals, closed-form lower and upper bounds on the average delay are characterized analytically. Secondly, for Poisson arrivals, a closed-form expression for the average delay is delineated using the Pollaczek-Khinchin formula. Based on the established bounds, the proposed approach achieves less average delay than single-layer outage approaches. Under optimal power allocation among the encoded layers, numerical evaluations demonstrate that the proposed approach significantly minimizes average sum delays compared to traditional outage approaches, especially under high arrival rates.

Keywords: broadcast approach; channel state information; latency; multiple access

check for updates

Citation: Zohdy, M.; Tajer, A.; Shamai, S. Broadcast Approach to Uplink NOMA: Queuing Delay Analysis. *Entropy* **2022**, *24*, 1757. https://doi.org/10.3390/e24121757

Academic Editors: Onur Günlü and Holger Boche

Received: 16 September 2022
Accepted: 15 November 2022
Published: 30 November 2022

1. Introduction

There is a growing need for maintaining low latency and high reliability in a wide range of wireless communication systems [1]. Among the recently proposed techniques for attaining the latency-reliability requirements is the power domain non-orthogonal multiple access (NOMA) [2–6]. Uplink power domain NOMA [5] facilitates simultaneous multi-user channel access, alleviating the traditional signaling period at the beginning of the transmission. Furthermore, by leveraging power control and adaptive decoding order among users, NOMA techniques enhance user fairness by taking into consideration the dissimilarities in the channel state of each user [7,8].

A fundamental challenge that NOMA faces in wireless networks is that its power control critically relies on the availability of full channel state information at each transmitter (CSIT). This assumption is generally unfeasible under the anticipated network scale growth. In the absence of CSIT, traditional NOMA occasionally suffers from outage events, which necessitate repeated re-transmissions and negatively affect the overall latency. To address this issue, we propose a non-orthogonal multi-access technique in which each transmitter splits its stream of information into multiple encoded layers, each adapted to a specific combination of all the network's channel states. Each user then transmits the superposition of all its encoded layers to the receiver. In particular, we approach the problem of minimizing the overall communication latency from a cross-layer resource allocation perspective by focusing on the dominant delay factor, i.e., the queuing delay [9]. The goal of the proposed approach is to minimize the average sum-queuing delay among users by optimally allocating power among the encoded layers at each transmitter in the physical layer.

Outage avoidance via multi-layer superposition coding was first proposed in [10,11] for the slowly fading single-user channels. This is generally referred to as the broadcast approach [12]. Furthermore, the studies in [13] extended the broadcast approach to the energy harvesting settings, those in [14–20] to random and multi-access channel models, and those in [21,22] to the multiuser interference channel. Aside from analyzing the achievable rate regions of multi-layer superposition coding [17,23], the average delay performance has only been studied for the single-user fading channel in [24]. However, under CSIT uncertainties, the advantages of adaptive multi-layer superposition coding for controlling the average queuing delay in multiple access channels are yet to be explored. Finally, we note that the broadcast approach is related to the studies on the "rate-splitting", the foundations of which rely on superposition coding of the layered information messages [25].

In this paper, we consider an N-user block fading multiple access channel (MAC) in which all transmitters are oblivious to their instantaneous channel state. Each user possesses an infinite capacity queue, occasionally holding the arriving information packets to be transmitted. A novel multi-layer superposition coding scheme is then employed, in which each transmitter adapts its message to the combined network state. Based on the proposed scheme, closed-form lower and upper bounds on the average delay are characterized analytically for deterministic arrivals. Furthermore, a closed-form expression for the average queuing delay is delineated for Poisson arrivals. Based on the derived bounds on average delay, the proposed approach is shown to outperform the single-layer outage approach. Finally, under optimal power allocation among the encoded layers, numerical evaluations demonstrate that the broadcast approach significantly reduces the average sum delays compared to traditional outage approaches under symmetric/asymmetric arrival rates and channel statistics among users.

A rich literature exists on minimizing the average delay through cross-layer resource allocation in MAC with full CSIT. Relevant studies include [26] in which the authors provide an optimal solution for minimizing average delays of two-user MAC channels by controlling the departure probability of each user's queue. In [27], an information-theoretic rate allocation policy is proposed to achieve a lower bound on the average delay of multi-access coding schemes. Dynamic power and rate control to minimize the average delay are studied for multi-access channels in [28]. The study in [28] provides a one-step value iteration policy for optimal scheduling in MAC fading channels. A lower bound on the LTE-A average delay is derived in [29] for random access channels under different arrival processes. The random access scheduling problem is addressed in [30] using a distributed virtual queue model facilitating a self-organizing policy. The study in [31] proposes a joint superposition coding and scheduling policy for the uplink NOMA by relying on user-pairing to reduce the complexity of analysis [32,33]. The accuracy of ranking users in NOMA techniques using distance-based measures versus instantaneous signal-to-noise ratio (SNR) is addressed in [34]. Joint scheduling and superposition coding in fading channels is studied in [35]. The effect of unsaturated traffic in uplink NOMA is studied in [36] using tools from queuing theory. Interaction between power control and queuing service rates in interference-limited channels is studied in [37]. Delay analysis of multi-point to multi-point networks is provided in [38] for spatial-temporal random arrival traffic. The problem of power control in delay-bounded applications is considered in [39], especially under the assumption of imperfect successive interference cancellation in uplink NOMA. The effective capacity of two-user uplink NOMA is characterized in [40] under quality-of-service delay constraints.

Energy-efficient transmission in uplink NOMA is studied in [41] under statistical delay constraints, where probabilistic upper bounds on queuing delays of NOMA are characterized. Resorting to the concept of effective capacity, the study in [42] proposes an optimized hybrid approach between non-orthogonal multiple access and orthogonal multiple access with different user pairing techniques in order to maximize the effective capacity under stringent delay constraints. Contention-based modified NOMA for uplink access is studied in [43], showing that exploiting collisions in the power domain can greatly reduce access delay. The throughput, access delay, and energy efficiency of NOMA uplink random access system are studied in [44]. Joint power control and user scheduling is considered in [45] to investigate the access delay minimization problem through an efficient sub-optimal iterative algorithm. Optimal power level partitioning to accommodate non-critical and high-priority messages is studied in [46]. A joint dynamic power control and user pairing algorithm is proposed in [47] to minimize long-term time average transmit power and queuing delay. Recent studies further includes [48] in which an adaptive rate NOMA with full CSIT is shown to provide better ergodic capacities for mobile users than OMA while satisfying strict local delay constraints for the internet of things (IoT) devices in cellular IoT networks. Opportunistic NOMA schemes are proposed in [49] for short message delivery with delay constraint based on which an upper bound on session error probability is derived, showing the impact of NOMA on session error under Rayleigh fading. A queuing delay analysis is presented in [50] for uplink NOMA with full CSIT, and the impact of channel estimation imperfections for finite-length channel coding is studied. Dynamic power allocation schemes with statistical delay quality-of-service (QoS) guarantees are shown in [51] to significantly improve the sum effective capacity and effective energy efficiency for an uplink NOMA system with paired users.

The rest of this paper is organized as follows. Section 2 presents the N-user multi-access channel model. The proposed multi-layer-based multi-access approach is outlined in Section 3 for the special case of the 2-state channel. The average delay achievable by the proposed approach is shown to outperform the average delay of the single-layer outage approach in Section 4 for deterministic and stochastic arrivals processes. The proposed multi-access approach is generalized to the case of finite arbitrary ℓ-state channel in Section 5. Finally, numerical evaluations are provided in Section 6, and the paper is concluded in Section 7.

2. Channel Model

Consider an N-user block fading MAC channel consisting of N transmitters and one receiver. The channel state is assumed to remain unchanged during the period of one transmission block of n channel uses and varies independently among consecutive blocks. We assume that the block length n is large enough to give rise to the notion of reliable communications but much shorter than the dynamics of the fading process [24]. Each transmitter is assumed to know the statistics of the channel state information (CSI) of its own link to the receiver but is oblivious to its instantaneous value. Complete CSI of all links is assumed to be available at the receiver. The input-output relationship of this channel is given by

$$Y = \sum_{i=1}^{N} h_i X_i + W\,, \tag{1}$$

where X_i denotes the transmitted signal from user i and W is the additive white Gaussian noise with zero mean and unit variance. Finally, h_i denotes the state of the fading channel between transmitter i and the receiver. The transmitted signal X_i is subject to an average power constraint P for all $i \in \{1, \ldots, N\}$, i.e., $\mathbb{E}\big[|X_i|^2\big] \leq P$. We consider a quantized model for the fading channel according to which h_i^2 takes one of two possible states, referred to as $\{\textit{weak}, \textit{strong}\}$, denoted by $\{\alpha_1, \alpha_2\}$, respectively. Without loss of generality, we assume $0 < \alpha_1 < \alpha_2 < +\infty$. User i experiences *strong* or *weak* channel states with probabilities $p_i \triangleq \mathbb{P}(h_i^2 = \alpha_2)$ and $\bar{p}_i \triangleq 1 - p_i$, respectively.

Each transmitter is assumed to possess an infinite-capacity queue. The queue at transmitter i receives random packets with an average arrival rate λ_i (bits/channel use). The size of the data queued at transmitter i at the beginning of any transmission block t is denoted by $\tilde{Q}_i(t)$, $\forall i \in \{1, \dots, N\}$. We define $A_i(t)$ as the total number of bits arriving in the queue at transmitter i during transmission block t. Finally, $r_i(t)$ (bits/channel use) denotes the service rate of the queue at transmitter i. Hence, the queue size at transmitter i at the end of any transmission block can be expressed using a recursive relationship as

$$\tilde{Q}_i(t+1) = \begin{cases} \tilde{Q}_i(t) + nA_i(t) - nr_i(t), & \tilde{Q}_i(t) + nA_i(t) - nr_i(t) \geq 0 \\ 0, & \text{otherwise} \end{cases}. \tag{2}$$

Accordingly, we define $Q_i(t)$ as queue size normalized by the number of transmission blocks n, i.e.,

$$Q_i(t+1) \triangleq \begin{cases} Q_i(t) + Z_i(t), & Q_i(t) + Z_i(t) \geq 0 \\ 0, & o.w. \end{cases}, \tag{3}$$

where the random variable $Z_i(t)$ is defined as $Z_i(t) \triangleq A_i(t) - r_i(t)$, and it captures the change in the queue size at transmitter i at the end of transmission block t. We remark that the number of bit arrivals $A_i(t)$ is random and does not necessarily fit into the exact size of the transmitted packet in a given transmission block. Therefore, if the backlogged data at any queue is less than a packet length, the data bits are zero-padded to form a complete packet for the encoder at each transmitter. Throughout the rest of the paper, we assume that the processing delay, i.e., encoding and decoding processes, as well as the transmission delay, are fixed and negligible with respect to the queuing delay. We use the concise notation $C(x, y) \triangleq \frac{1}{2}\log_2(1 + \frac{x}{\frac{1}{P}+y})$, $\{x^i_j\}^k_{j=1} \triangleq \{x^i_1, x^i_2, \dots, x^i_k\}$. Finally, we denote the set of all users in the network by $\mathcal{N} \triangleq \{1, \dots, N\}$.

3. 2-State Channel Multi-Access

In this section, we present a non-orthogonal multiple-access approach based on multi-layer encoding at each transmitter and successive interference cancellation (SIC) at the receiver. The underlying layering approach hinges on adapting the number of encoded layers at each transmitter to the combined fading state of the network, i.e., the fading states of all transmitters to the receiver. Owing to the arising interference in non-orthogonal multi-access channels with no CSIT, the channel state of each user directly affects the decoding success probabilities of all the other users. Motivated by this, the recent work in [17] proposed a multi-layer coding approach for the two-user multiple access channel with no CSIT, specially adapted to the combined network state resulting in an enlarged average achievable rate regions compared to the existing multi-layer coding approaches. In this section, we extend the layering approach in [17] to the general case of an arbitrary number of N-users. As shown in this paper, the proposed multi-access approach enjoys considerable advantages in reducing the queuing delay.

3.1. Layering Approach

At the beginning of each transmission block, user i aims to transmit all the data bits accumulated in its queue if the channel state allows it. Otherwise, it encodes a part of its data with the maximum allowable encoding rate. Towards this goal, user i encodes its data (fully or partially) using $2N$ independent messages generated from $2N$ Gaussian codebooks. These messages are denoted by $U^i_{jk}, \forall i \in \mathcal{N}, j \in \{1, 2\}, k \in \{0 \cup \mathcal{N}\}$. Based on this decomposition

$$X_i = \sum_{j=1}^{2} \sum_{k=0}^{N} U^i_{jk} \,. \tag{4}$$

We consider an ordering of the network states based on the number of users with *strong* channel states denoted by k. We define $\mathcal{S}_k$ as the set of k users' indices that experience *strong* channel states. Accordingly, $\mathcal{E}_k$ denotes the event that exactly k users are experiencing a *strong* channel including.

The notation U^i_{jk} can be interpreted as follows. Superscript i denotes the user index $i \in \mathcal{N}$, subscript $j \in \{1, 2\}$ refers to user i's channel state, where $j = 1$ if $h_i^2 = \alpha_1$ and otherwise $j = 2$. Finally, $k \in \{0 \cup \mathcal{N}\}$ represents the number of users in the network with a *strong* channel state, possibly including user i's channel. Therefore, for every value of k, user i adapts the rate of two codewords, $\{U^i_{jk}\}_{j=1}^2$, based on its own channel state resulting in a total of $2k$ layers. The correspondence between each channel state and the adapted layer is shown in Table 1 and summarized below:

- U^i_{10} is adapted to $\mathcal{E}_0$, where all channels are *weak*.
- U^i_{2N} is adapted to $\mathcal{E}_N$, where all channels are *strong*.
- When exactly k channels are *strong*:
  - U^i_{1k} is adapted to $\mathcal{N} \backslash \mathcal{E}_k$ if user i's channel is *weak*.
  - U^i_{2k} is adapted to $\mathcal{E}_k$ if user i's channel is *strong*.

The rate of codeword U^i_{jk} is denoted by R^i_{jk}. Finally, we define β^i_{jk} as the power fraction of the total power P allocated to codeword U^i_{jk}, such that

$$\sum_{j=1}^{2} \sum_{k=0}^{N} \beta^i_{jk} = 1 \ .$$

For user i, the rate of each codebook is governed via the power allocation parameters β^i_{jk} such that at least one layer is successfully decoded in every possible network state.

Table 1. Layering and codebook assignments by user i.

h_i^2 \ k	**0**	**1**	**2**	...	$N-1$	N
α_1	U^i_{10}	U^i_{11}	U^i_{12}	...	U^1_{1N-1}	
α_2		U^i_{21}	U^i_{22}	...	U^i_{2N-1}	U^i_{2N}

3.2. Decoding Approach

Corresponding to the layering approach in Section 3.1, we propose a decoding algorithm with $2kN$ SIC stages for each combined channel state with k strong channels. The layers' decoding order is adapted to the combined channel states such that all the layers adapted to channel states with less than k strong users, $\{U^i_{j\ell}, \forall j \in \{1, 2\}, \ell < k\}$, are first decoded and subtracted from the received signal. Afterwards, layers adapted to channel state with exactly k strong users, $\{U^i_{jk}, \forall j \in \{1, 2\}\}$, are decoded.

When $|\mathcal{S}| = k$, the receiver employs $4k + 1$ decoding stages. Each of the layers for any $j \in \{1, 2\}$ and $\ell \in \{0, \ldots, k\}$, the set of codebooks $\{U^i_{j\ell} : i \in \mathcal{N}\}$ is partitioned to two sets

$$\mathcal{P}_{j\ell} \triangleq \{U^i_{j\ell} : i \in \mathcal{S}\} \qquad \text{and} \qquad \mathcal{Q}_{j\ell} \triangleq \{U^i_{j\ell} : i \notin \mathcal{S}\} \ , \tag{5}$$

rendering a total of $4k + 1$ partitions for different $j \in \{1, 2\}$ and $\ell \in \{0, \ldots, k\}$. The decoding strategy decodes one message from each of these, except for the partition $\{U^i_{2k} : i \notin \mathcal{S}\}$. The decoding strategy works as follows. We create the following two sequences of sets:

$$\mathcal{P} \triangleq \{\mathcal{P}_{10}, \mathcal{P}_{11}, \mathcal{P}_{21}, \ldots, \mathcal{P}_{2(k-1)}, \mathcal{P}_{1k}, \} \ , \tag{6}$$

$$\mathcal{Q} \triangleq \{\mathcal{Q}_{1k}, \mathcal{Q}_{2(k-1)}, \mathcal{Q}_{1(k-1)}, \ldots, \mathcal{Q}_{11}, \mathcal{Q}_{10}, \} \ . \tag{7}$$

The decoding strategy selects codebooks by alternating between $\mathcal{P}$ and $\mathcal{Q}$ in ascending order and decodes exactly one codebook from each. Specifically, the codebook sets are selected in the following order: $\{\mathcal{P}_{10}, \mathcal{Q}_{1k}, \mathcal{P}_{11}, \mathcal{Q}_{2(k-1)}, \mathcal{P}_{21}, \ldots, \mathcal{P}_{1k}, \mathcal{Q}_{10}\}$. This results in $4k$ coding stages. Finally, the codebooks in $\{U^i_{2k} : i \in \mathcal{S}\}$ are decoded as the last stage, i.e., stage $4k+1$. Next, we describe the decoding stages and the set of codebooks decoded in each.

- **Decoding stage** 1**:** We start by decoding the layers $\mathcal{P}_{10} \triangleq \{U^i_{10} : i \in \mathcal{S}\}$, i.e., the codebooks U^i_{10} of only the k strong users in $\mathcal{S}$. We define $\mathcal{S}_k$ as an ordered set of these users, in which the users are ordered in an ascending order based on their indices. The codebooks will be decoded sequentially in this order.
- **Decoding stage** 2**:** Next, after decoding and removing the codebooks in $\mathcal{P}_{10}$, we sequentially decode the layers in $\mathcal{Q}_{1k} = \{U^i_{1k} : i \notin \mathcal{S}\}$, which involves layers U^i_{1k} of users with weak channels.
- **Decoding stage** 3**:** In the third stage, the codebooks in $\mathcal{P}_{10}$ and $\mathcal{Q}_{1k}$ are already decoded. We continue by sequentially decoding the set of codebooks in $\mathcal{P}_{11} \triangleq \{U^i_{11} : i \in \mathcal{S}\}$.
- **Decoding stage** 4**:** The decoding process continues by sequentially decoding the codebooks in $\mathcal{Q}_{2(k-1)} = \{U^i_{2(k-1)} : i \notin \mathcal{S}\}$, while the codebooks of $\mathcal{P}_{10}$, $\mathcal{Q}_{1k}$, and $\mathcal{P}_{11}$ are already decoded.
- **Decoding stage** 5**:** This stage sequentially decodes the codebooks $\mathcal{P}_{21}$.
- **Decoding stage** 6**:** This stage sequentially decodes the codebooks in $\mathcal{Q}_{1(k-1)}$.
- **Decoding stages** $\{2,\ldots,4k+1\}$**:** Following the pattern of the previous decoding stages, in general, in stage $\{2,\ldots,4k\}$, we decode the codebooks according to the following schedule for $\ell \in \{1,\ldots,k\}$:

$$\begin{array}{lll} \text{codebooks in } \mathcal{Q}_{1(k-\ell+1)} & \text{stage} & 4\ell-2 \\ \text{codebooks in } \mathcal{P}_{1\ell} & \text{stage} & 4\ell-1 \\ \text{codebooks in } \mathcal{Q}_{2(k-\ell)} & \text{stage} & 4\ell \\ \text{codebooks in } \mathcal{P}_{2\ell} & \text{stage} & 4\ell+1 \end{array} \tag{8}$$

The proposed decoding approach results in decoding more layers for a channel state with k strong users compared to a state with $k-1$ strong users. In particular, the receiver decodes one extra layer for user i in channel state $\mathcal{E}_k$ as compared to state $\mathcal{E}_{k-1}$. Note that in both states, user i experiences a *weak* channel. On the other hand, the receiver decodes two extra layers for user i in channel state $\mathcal{E}_k$ as compared to state $\mathcal{E}_{k-1}$, note that user i experiences a *strong* channel in both states. Our intuition behind such a strategy hinges on two factors. First, that decoding and removing additional interfering users with strong channel states is expected to increase the achievable rate of user i. Secondly, when user i experiences a stronger channel, the receiver can possibly decode an additional layer from its message. The decoded layers for channel state $\mathcal{E}_k$ are shown in Table 2 for illustration.

Table 2. Decoded layers for channel state $\mathcal{E}_k$ where $h_i^2 = \alpha_j$.

Stage	**Stage 1**	**Stage 2**	**Stage 3**	**Stage 4**	...	**Stage 4$k+1$**
Codebook	$\mathcal{P}_{10}$	$\mathcal{Q}_{1k}$	$\mathcal{P}_{11}$	$\mathcal{Q}_{2(k-1)}$	...	$\{U^1_{2k} : i \in \mathcal{S}_k\}$

Finally, the detailed steps of the proposed successive decoding algorithm are presented in Algorithm 1. We remark that the effect of the precedence of users with similar channel states within each decoding stage on the average achievable delay will be analyzed in the subsequent sections.

Algorithm 1: Successive Decoding for 2-state channel

```
input $(h_1^2, \ldots, h_N^2), k$
for $\ell \in \{0, \ldots, k\}$
  if $\ell = 0$
    In stage 1 successively decode $\{U^i_{10}\}_{i=1}^N$
  else if $\ell \in \{1, \ldots, k\}$
    (1) In stage $4\ell - 2$ successively decode $\mathcal{Q}_{1(k-\ell+1)}$
    (2) In stage $4\ell - 1$ successively decode $\mathcal{P}_{1\ell}$
    (3) In stage $4\ell$ successively decode $\mathcal{Q}_{2(k-\ell)}$
    (4) In stage $4\ell + 1$ successively decode $\mathcal{P}_{2\ell}$
  end if
end for
```

Based on the multi-access approach outlined throughout this section, the service rate of the queue at transmitter i is determined by the total rates of the successfully decoded layers during each network state. Therefore, the service rate $r_i(t)$ during transmission block t varies randomly and is jointly determined by the states of all users as well as the power allocation among different layers at each transmitter, i.e., β^i_{jk}. The achievable rates for all the encoded layers are formally stated in the Theorem 1.

Theorem 1. *For the N-user MAC channel without CSIT, when exactly $k \in \mathcal{N} \cup \{0\}$ users have strong channels, the achievable rates of the layering approach in Section 3.1 and the decoding policy in Algorithm 1 are characterized by the set of rates $\left\{R^i_{jk}, \forall j \in \{1,2\}, i \in \mathcal{N}, \ell \in \{0 \cup \mathcal{N}\}\right\}$ that satisfy*

$$R^i_{j\ell} \leq \min_{\mathcal{S}:|\mathcal{S}|=k} d^i_{j\ell}(\mathcal{S}) \,, \tag{9}$$

where constants $\left\{d^i_{jk}(\mathcal{S}), \forall k \in \{0 \cup \mathcal{N}\}, j \in \{1,2\}\right\}$ are defined in Appendix A.

Proof. See Appendix B. □

We remark that characterizing the achievable rate region of the proposed approach in the form of rate bounds on individual codebooks rates, rather than an average achievable rate region, will be instrumental to characterizing the average achievable delay analysis throughout the next section.

4. Average Queuing Delay

In this section, we investigate the average queuing delay achieved by the multi-access approach in Section 3 compared to the conventional single-layer (outage) multi-access approach. First, in Section 4.1, we focus on the case of the deterministic arrival process at each queue, for which we delineate lower and upper bounds on the average queuing delay. Furthermore, the case of stochastic arrivals is examined in Section 4.2 in which a closed-form expression for the average delay achievable by the proposed approach is characterized and compared to that of the single-layer transmission approach. To proceed, we define $\mathcal{E}^i_k$ as the event in which we have exactly k strong channels and they include the channel of user i. Accordingly, we define $\bar{\mathcal{E}}^i_k \triangleq \mathcal{N} \backslash \mathcal{E}^i_k$. We begin by computing the probabilities of the events $\mathcal{E}^i_k$ (and $\mathcal{E}(\bar{\mathcal{S}}^i_k)$) as follows.

$$\mathbb{P}\left[\mathcal{E}^i_k\right] = \sum_{\substack{\mathcal{I}\subseteq\mathcal{N}\\ |\mathcal{I}|=k}} \prod_{j\in\mathcal{I}} p_j \prod_{\substack{\ell\notin\mathcal{I}\\ \ell\neq i}} \bar{p}_\ell \quad \text{and} \quad \mathbb{P}\left[\bar{\mathcal{E}}^i_k\right] = \sum_{\substack{\mathcal{I}\subseteq\mathcal{N}\\ |\mathcal{I}|=k}} \prod_{\substack{j\in\mathcal{I}\\ j\neq i}} p_j \prod_{\ell\notin\mathcal{I}} \bar{p}_\ell \,. \tag{10}$$

where $\mathcal{I}$ denotes a subset of user indices.

4.1. Deterministic Arrivals

Throughout this subsection, we assume that the data arrival process at each queue is a deterministic process with an average arrival rate λ_i, i.e., $A_i(t) = \lambda_i$, $\forall i \in \mathcal{N}$. Note that as a result of the zero-padding applied by the encoder, whenever the available data bits are fewer than a transmission packet, a G/G/1 queuing model is generated at each transmitter. A closed-form expression characterizing the average delay of the G/G/1 queuing model is, in general, unknown. Therefore we resort to characterizing upper and lower bounds on the average queuing delay. These bounds are formally presented in Theorem 2. Before stating Theorem 2, we provide an outline of the main steps pertinent to deriving the characterized bounds, where the detailed proof can be found in Appendix C.

Establishing the desired bounds hinges on characterizing the average queue size at each transmitter i using the Laplace transform of the probability distribution function (PDF) of the queue size Q_i (moment generating function). Let the PDF of Q_i be denoted by $dF_i(q)$ and its associated Laplace transform be denoted by $L_i(s)$. Therefore, the average queue size at transmitter i is given by

$$\mathbb{E}[Q_i] = \lim_{s\to 0} -\frac{dL_i(s)}{ds} \,. \tag{11}$$

Recalling the recursive expression for Q_i in terms of the variable Z_i in (3), a recursive form of $F_i(q)$ can be expressed as follows [52,53]

$$F_i(q) = \begin{cases} \int_{-\infty}^{q} F_i(q-\tau) dF_{Z_i}(\tau) \,, & q \geq 0 \\ 0 \,, & q < 0 \end{cases} , \tag{12}$$

where $dF_{Z_i}(z)$ denote PDF of Z_i denoting change in queue size at user i. At the end of every transmission block, the change in queue size i, Z_i, is primarily determined by the difference between the data arrival λ_i and the total rate of all the layers successfully decoded by the receiver from user i's message stream, which in turn is determined by the combined network state. Consequently, $dF_{Z_i}(z)$ can be expressed as

$$\begin{aligned} dF_{Z_i}(z) = {} & \mathbb{P}\Big[\mathcal{E}\Big(\bar{\mathcal{S}}_0^i\Big)\Big]\delta\Big(z - \lambda_i + R_{10}^i\Big) + \mathbb{P}\Big[\mathcal{E}\Big(\mathcal{S}_N^i\Big)\Big]\delta\left(z - \lambda_i + \sum_{j=1}^{2}\sum_{k=1}^{N} R_{jk}^i\right) \\ & + \sum_{\ell=1}^{N-1} \mathbb{P}\Big[\mathcal{E}\Big(\mathcal{S}_\ell^i\Big)\Big]\delta\left(z - \lambda_i + \sum_{j=1}^{2}\sum_{k=0}^{\ell-1} R_{jk}^i + R_{2\ell}^i\right) \\ & + \sum_{\ell=1}^{N-1} \mathbb{P}\Big[\mathcal{E}\Big(\bar{\mathcal{S}}_\ell^i\Big)\Big]\delta\left(z - \lambda_i + \sum_{j=1}^{2}\sum_{k=0}^{\ell-1} R_{jk}^i + R_{1\ell}^i\right) . \end{aligned} \tag{13}$$

We remark that in order to guarantee the stability of the data queue at each transmitter, we assume that the arrival rate λ_i is less than the average achievable rate (service rate of the queue), i.e.,

$$\lambda_i < \mathbb{E}[r_i] \,, \qquad \forall i \in \mathcal{N} \,, \tag{14}$$

where the average service rate at queue i is given by

$$\begin{aligned} \mathbb{E}[r_i] = {} & \mathbb{P}\Big[\mathcal{E}\Big(\bar{\mathcal{S}}_0^i\Big)\Big] R_{10}^i + \mathbb{P}\Big[\mathcal{E}\Big(\mathcal{S}_N^i\Big)\Big] \cdot \sum_{j=1}^{2}\sum_{k=1}^{N} R_{jk}^i \\ & + \sum_{\ell=1}^{N-1} \mathbb{P}\Big[\mathcal{E}\Big(\mathcal{S}_\ell^i\Big)\Big] \cdot \left(\sum_{j=1}^{2}\sum_{k=0}^{\ell-1} R_{jk}^i + R_{2\ell}^i\right) + \sum_{\ell=1}^{N-1} \mathbb{P}\Big[\mathcal{E}\Big(\bar{\mathcal{S}}_\ell^i\Big)\Big] \cdot \left(\sum_{j=1}^{2}\sum_{k=0}^{\ell-1} R_{jk}^i + R_{1\ell}^i\right) . \end{aligned} \tag{15}$$

An explicit expression for $F_i(q), \forall i \in \mathcal{N}$, directly follows by combining (12) and (13)

$$F_i(q) = \begin{cases} 0\,, & \forall q \in \mathcal{R}_1 \\ \mathbb{P}[\mathcal{E}(\mathcal{S}_0^i)]\,F_i(q - \lambda_i + \sum_{j=1}^{2}\sum_{k=0}^{N} R_{jk}^i)\,, & \forall q \in \mathcal{R}_2 \\ \quad\vdots & \\ \mathbb{P}[\mathcal{E}(\bar{\mathcal{S}}_0^i)]\,F_i(q - \lambda_i + R_{10}^i)\,, & \forall q \in \mathcal{R}_{2N-1} \end{cases}\,, \tag{16}$$

where the intervals $\mathcal{R}_i, \forall i \in \{1, \dots, 2N-1\}$, are given by

$$\mathcal{R}_1 \triangleq (-\infty, 0)\,,$$
$$\mathcal{R}_2 \triangleq \left[0, \lambda_i - \sum_{j=1}^{2}\sum_{k=0}^{N} R_{jk}^i + R_{1(N-1)}^i\right]\,,$$
$$\vdots$$
$$\mathcal{R}_{2N-1} \triangleq \left[\lambda_i - R_{10}^i, \infty\right)\,.$$

Finally, the Laplace transform of the queue size PDF is computed using (16), which in turn facilitates obtaining the average queue size at user i. Note that although $F_i(q)$ is expressed in (16), it is still a recursive form. Therefore, the obtained expression for the average queue size delay contains the unknown term $F_i(q)$, which is why a closed form cannot be obtained. Subsequently, an upper and a lower bound on the average queue size of user $i \in \mathcal{N}$ are formally characterized in the next theorem.

Theorem 2. *The average queue size of transmitter i under the multi-access policy in Section 3 is bounded by*

$$\frac{1}{2}\sum_{j=1}^{2}\sum_{k=0}^{N} R_{jk}^i - \frac{\lambda_i}{2} - \frac{N_i}{D_i} \leq \mathbb{E}[Q_i] \leq \sum_{j=1}^{2}\sum_{k=0}^{N} R_{jk}^i - \lambda_i - \frac{N_i}{D_i}\,, \tag{17}$$

where we have defined $D_i \triangleq 2(\mathbb{E}[r_i] - \lambda_i)$ *and*

$$N_i \triangleq -\left(\sum_{j=1}^{2}\sum_{k=0}^{N} R_{jk}^i - \lambda_i\right)^2 + \mathbb{P}\left[\mathcal{E}\left(\bar{\mathcal{S}}_0^i\right)\right]\left(\sum_{j=1}^{2}\sum_{k=1}^{N} R_{jk}^i\right)^2$$
$$+ \sum_{\ell=1}^{N-1}\mathbb{P}\left[\mathcal{E}\left(\mathcal{S}_\ell^i\right)\right]\cdot\left(\sum_{j=1}^{2}\sum_{k=\ell+1}^{N} R_{jk}^i + R_{1\ell}^i\right)^2 + \sum_{\ell=1}^{N-1}\mathbb{P}\left[\mathcal{E}\left(\bar{\mathcal{S}}^i{}_\ell\right)\right]\cdot\left(\sum_{j=1}^{2}\sum_{k=\ell+1}^{N} R_{jk}^i + R_{2\ell}^i\right)^2. \tag{18}$$

Proof. See Appendix C. □

Using Little's law, upper and lower bounds on the average queuing delay at transmitter i under deterministic arrivals can directly be obtained by normalizing the bounds characterized in Theorem 2 $\mathbb{E}[Q_i]$ by λ_i.

In order to assess the performance of the proposed multi-layer superposition coding access approach, we compare the achievable average queuing delay to that of the conventional single-layer access (outage) approach. To this end, we first summarize the single-layer approach, and afterward, a lower bound on the average queuing delay achieved by the single-layer approach is characterized in Lemma 1. Finally, we compare the rate of increase of the average delay achieved by each policy with respect to the data arrival rate. As the arrival rate increases, the rate of increase of the average delay with respect to λ_i resulting from the proposed approach is lower than that resulting from the single-layer (outage) approach.

According to the single-layer (outage) transmission approach, each transmitter encodes the available data in its queue into one layer of a fixed rate irrespective of the unknown network state. For $i \in \mathcal{N}$, let $R_i^{\rm s}$ denote the rate of the single encoded layer transmitted by user i in the outage approach. In any given transmission block, if the rate R_i^{s} lies in the achievable rate region of the actual network state, it will be successively decoded by the receiver. Otherwise, an outage occurs where the receiver fails to decode the message of user i, and the transmitter attempts to re-transmit the same message in the subsequent transmission block using the same encoding rate R_i^{s}. We define $r_i^{\mathrm{s}}(t)$ as the service rate of the queue at user i under the single-layer transmission, the encoding rate of the codeword transmitted by user i in transmission block t and successively decoded by the receiver, hence removed from user i's queue. Furthermore, we denote by p_i^{s} the probability of successfully decoding a message of rate R_i^{s} from user i. Accordingly, the service rate of the queue at transmitter i using the outage approach is given by

$$r_i^{\mathrm{s}}(t) = \begin{cases} R_i^{\mathrm{s}}, & \text{with probability} \quad p_i^{\mathrm{s}} \\ 0, & \text{with probability} \quad 1 - p_i^{\mathrm{s}} \end{cases}. \tag{19}$$

Finally, we define Q_i^{s} as the queuing size at transmitter i under the single-layer transmission approach summarized above. In Lemma 1, we characterize lower and upper bounds on the average $\mathbb{E}[Q_i^{\mathrm{s}}]$ using an approach similar to that used to characterize the bounds in Theorem 2.

Lemma 1. *The average queue size of transmitter i under single layer (outage) approach is lower and upper bounded according to:*

$$\frac{1}{2}R_i^{\mathrm{s}} - \frac{\lambda_i}{2} - \frac{\left(R_i^{\mathrm{s}} - \lambda_i\right)^2 - R_i^{\mathrm{s}}\left(1 - p_i^{\mathrm{s}}\right)}{2\left(p_i^{\mathrm{s}}R_i^{\mathrm{s}} - \lambda_i\right)} \leq \mathbb{E}[Q_i^{\mathrm{s}}] \leq R_i^{\mathrm{s}} - \lambda_i - \frac{\left(R_i^{\mathrm{s}} - \lambda_i\right)^2 - R_i^{\mathrm{s}}\left(1 - p_i^{\mathrm{s}}\right)}{2\left(p_i^{\mathrm{s}}R_i^{\mathrm{s}} - \lambda_i\right)}. \tag{20}$$

Proof. Follows the same argument as that in Appendix C. □

In Theorem 2 and Lemma 1, we remark that the characterized bounds on the average queuing delay at each transmitter depend only on the arrival rate at the same node. Therefore, the effect of the average arrival rate on the delay bounds in (17) or (20) can be analyzed for each node i independently. In Theorem 3, while fixing the average achievable rates at each user among both approaches, we show that as the arrival rate λ_i at each user increases, the proposed multi-access approach lower rate of increase in the average queuing delay with respect to that achieved by the single layer approach.

Theorem 3. *For the N-user multiple access channel, given that*

$$\mathbb{E}[r_i] = \mathbb{E}[r_i^{\mathrm{s}}], \tag{21}$$

the rate of increase of average delay with respect the arrival rate under the approach in Section 3 is lower than that achieved by single-layer outage approach, i.e., for every $i \in \mathcal{N}$

$$\frac{\partial \mathbb{E}[Q_i]}{\partial \lambda_i} \leq \frac{\partial \mathbb{E}[Q_i^{\mathrm{s}}]}{\partial \lambda_i}. \tag{22}$$

Proof. See Appendix D. □

4.2. Stochastic Arrivals

In this section, we consider the proposed multi-layer superposition coding policy presented in Section 3 under Poisson distributed random arrivals $A_i \sim Pois(\lambda_i)$. We adopt the same queuing model in which each transmitter applies zero-padding in case the available bits in its queue are fewer than the size of a transmitted packet. Therefore, under Poisson distributed arrivals, the considered model constitutes an $M/G/1$ queuing model with an average arrival rate λ_i and service rate r_i specified in (15). Furthermore, we denote the queue utilization at transmitter i by $\rho_i \triangleq \frac{\lambda_i}{\mathbb{E}[r_i]}$. The average queue length for an $M/G/1$

queue can be characterized in a closed form by directly applying the Pollaczek-Khinchin formula. Theorem 4 formally states the average queuing size under the proposed layering and decoding approach.

Theorem 4. *According to the multi-access approach outlined in Section 3, the average queue length at user i with Poisson distributed arrivals with the average rate λ_i is given by*

$$\mathbb{E}[Q_i] = \rho_i + \frac{\rho_i^2 + \lambda_i \mathbb{V}[r_i]}{2(1-\rho_i)}\,, \tag{23}$$

where the average service rate $\mathbb{E}[r_i]$ is given by (15) and the variance of the service rate $\mathbb{V}[r_i]$ is

$$\begin{aligned}\mathbb{V}[r_i] = -\mathbb{E}[r_i] + \mathbb{P}\left[\mathcal{E}\left(\bar{\mathcal{S}}_0^i\right)\right](R_{10}^i)^2 + \mathbb{P}\left[\mathcal{E}\left(\mathcal{S}_N^i\right)\right] \cdot \left(\sum_{j=1}^{2}\sum_{k=1}^{N} R_{jk}^i\right)^2 \\ + \sum_{\ell=1}^{N-1} \mathbb{P}\left[\mathcal{E}\left(\mathcal{S}_\ell^i\right)\right] \cdot \left(\sum_{j=1}^{2}\sum_{k=0}^{\ell-1} R_{jk}^i + R_{2\ell}^i\right)^2 + \sum_{\ell=1}^{N-1} \mathbb{P}\left[\mathcal{E}\left(\bar{\mathcal{S}^i}_\ell\right)\right] \cdot \left(\sum_{j=1}^{2}\sum_{k=0}^{\ell-1} R_{jk}^i + R_{1\ell}^i\right)^2.\end{aligned} \tag{24}$$

Proof. Follows by applying Pollaczek-Khinchin formula for the M/G/1 average queue size [54], where the service rate of queue i is given by r_i. □

We remark that the proof of Theorem 3 implies that the proposed approach outperforms the single-layer outage approach in the case of Poisson arrivals as well, under equal average achievable rates. This result can be readily verified given that the proof in Appendix D essentially boils down to showing that the variance of the service rate (transmission rate) at each queue, $\mathbb{V}[r_i]$, is higher in the case of single-layer outage approach when compared to the proposed multi-layer approach.

5. ℓ-State Channel Multi-Access

In this section, we generalize the multi-access encoding and decoding approach outlined in Section 3 from the special case of 2-state channel, $\{$*weak*, *strong*$\}$, to channel with an arbitrary number of states ℓ. We denote the channel states by $\{\alpha_1, \ldots, \alpha_\ell\}$. Without loss of generality, we assume that $0 < \alpha_1 < \cdots < \alpha_\ell < +\infty$. Similarly to Section 2, we consider a slowly fading non-orthogonal multiple access channel model with N-transmitters and one receiver. The channel power gain of each user i can randomly take one of ℓ-states, i.e., $h_i^2 \in \{\alpha_1, \ldots, \alpha_\ell\}$.

In the layering approach in Section 3.1, we ordered the network state according to the number of users experiencing a *strong* channel state. Subsequently, each user splits its message into $2N$ layers, and the receiver decodes the layers adapted to the actual network state. Similarly, for the ℓ-state channel, we order the combined network state according to the number of users in the network sharing a particular state α_j as well as the value of such a state. In particular, a combined network state is degraded with respect to another state if it has a strictly smaller sum-rate capacity. We define the column vector $\boldsymbol{h} \overset{\triangle}{=} [h_1^2, \ldots, h_N^2]^T$ as the the combined network state and consider that a network state $\boldsymbol{h}$ to be degraded with respect to network state $\tilde{\boldsymbol{h}}$ if and only if

$$\|\boldsymbol{h}\|_1 < \|\tilde{\boldsymbol{h}}\|_1\,. \tag{25}$$

The motivation of such ordering stems from the fact the condition in (25) indicates the state $\tilde{\boldsymbol{h}}$ allows higher sum-rate capacity in an N-user MAC with full CSIT. In order to overcome the absence of full CSIT at each user, a transmitter splits its message into a finite number of layers, each adapted to the combined network state to avoid complete outages. Similarly to Section 3.1, user i encodes an available message using $(\ell-1)N+1$ independent random Gaussian codebooks. The codewords of these codebooks are denoted by U_{jk}^i. For layer U_{jk}^i, $j \in \{1, \ldots, \ell\}$ denotes the channel state of user i, that is $h_i^2 = \alpha_j$, while

$k \in \{0, \ldots, N-1\}$ denotes the number of users in the network with stronger channel state, i.e., $k = \sum_{i=1}^{N} I(h_i^2 > \alpha_j)$ where $I(x)$ is the indicator function.

According to the layering approach outlined above, the receiver attempts to successively decode up to $N((\ell-1)N+1)$ depending on the exact combined network state $\boldsymbol{h}$. In particular, when the actual network state is $\boldsymbol{h}$, the receiver decodes for each user i layer U_{jk}^i adapted to network state $\boldsymbol{h}$ in addition to all the layers adapted to all degraded network states $\tilde{\boldsymbol{h}}$ such that (25) is satisfied. The number of layers decoded for user i at the receiver increases from network state $\boldsymbol{h}$ to network state $\hat{\boldsymbol{h}}$ either if its own channel state becomes stronger or if the number of users experiencing channels strictly stronger than h_i^2 increases.

Given a network state $\boldsymbol{h}$, the receiver employs up to M stages of successive decoding, where M denotes the argument of the strongest channel gain in the network, i.e., $M \stackrel{\triangle}{=} \arg \parallel \boldsymbol{h} \parallel_\infty$. In stage $n \in \{1, \ldots, M\}$, the receiver successively decodes up to one layer for each user according to a descending order of the channel states among users. The details of the proposed decoding order for the ℓ-state channel are outlined in Algorithm 2.

Algorithm 2: Successive Decoding for ℓ-state channel

```
input $\boldsymbol{h}$
set $k_i = \sum_{d=1}^{N} I(h_d^2 > \alpha_i), \forall i \in \mathcal{N}, \quad M \stackrel{\triangle}{=} \arg \parallel \boldsymbol{h} \parallel_\infty$
for $m \in \{1, \ldots, M\}$
    Successively decode $\{U_{mk_i}^i : h_i^2 \geq \alpha_m, \ \forall i \in \mathcal{N}\}$ .
end for
```

We remark that according to the proposed layering approach for the ℓ-state channel and decoding approach in Algorithm 2, the total number of layers decoded by the receiver from each user i is possibly different in certain network states. Although, one possible generalization of the layering policy in Section 3 is that each user adapts a different encoding layer to each possible combined channel state, which in turn requires each user to encode its message into ℓ^N layers. However, the computational complexity of the decoding process, in addition to determining the optimal power allocation among layers, is considerable as the number of users N grows larger. Therefore, we adopt the outlined layering approach where each user splits its message into $N(\ell-1)+1$ layers instead of ℓ^N layers.

6. Numerical Evaluations

In this section, we evaluate the average achievable queuing delay for each user in the MAC channel using the multi-access broadcast approach outlined in Section 3. In particular, we adopt a Monte-Carlo simulation to optimally allocate the transmission power among the encoded layers at each user such that the average queuing delay is minimized. We divide the comparison settings into two main parts according to the arrival process at each queue, where we set the arrival process to be the same among both users in each setting. The first considers deterministic arrivals with value λ. The second one considers the Poisson arrival process. Furthermore, we also consider symmetric and asymmetric channel distributions among users. Throughout this section, we set the channel gains to $\alpha_1 = 0.5$ for the weak channel and $\alpha_2 = 1$ for the strong channel gains. In the symmetric case, we set the channel probability distribution for each user as $p_1 = p_2 = 0.5$, and in the asymmetric case, we set the probabilities to $p_1 = 0.5$ and $p_2 = 0.1$. In the asymmetric model, user 2 encounters a weak channel with a high probability, i.e., $\bar{p}_2 = 0.9$. We set the objective function in this numerical simulation to minimize the sum average delays of users 1 and 2 for the broadcast approach. Subsequently, based on the obtained optimal power distribution among the layers at each user, we evaluate the resulting average delay for the outage approach such that the average rates for each user are equal across both approaches.

Figures 1 and 2 focus on deterministic arrivals in the symmetric and asymmetric channel settings. In these figures, we compare average delay versus varying arrival rate λ in the proposed broadcast approach (denoted by "Bc") and in the outage approach (denoted by "outage"). In these evaluations, we have set the SNR to $P = 10$ dB. Furthermore, in these figures, we provide upper bounds that we have characterized for the broadcast approach (denoted by "Bc_{UB}") and the outage approach (denoted by "$\text{Outage}_{\text{UB}}$"). Figures 3 and 4 depict the counterparts of these results for Poisson arrival processes. Finally, it is observed that introducing asymmetry in the models (i.e., unequal probabilities for encountering strong channels) slightly improves the average latency of the broadcast approach, whereas it does not have a notable effect in the outage approach.

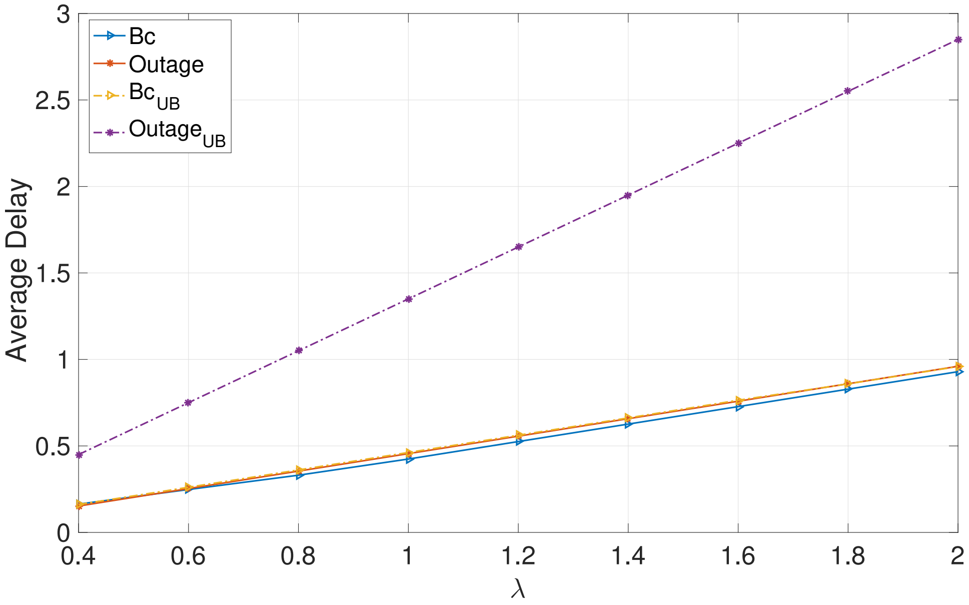

Figure 1. Deterministic: Symmetric.

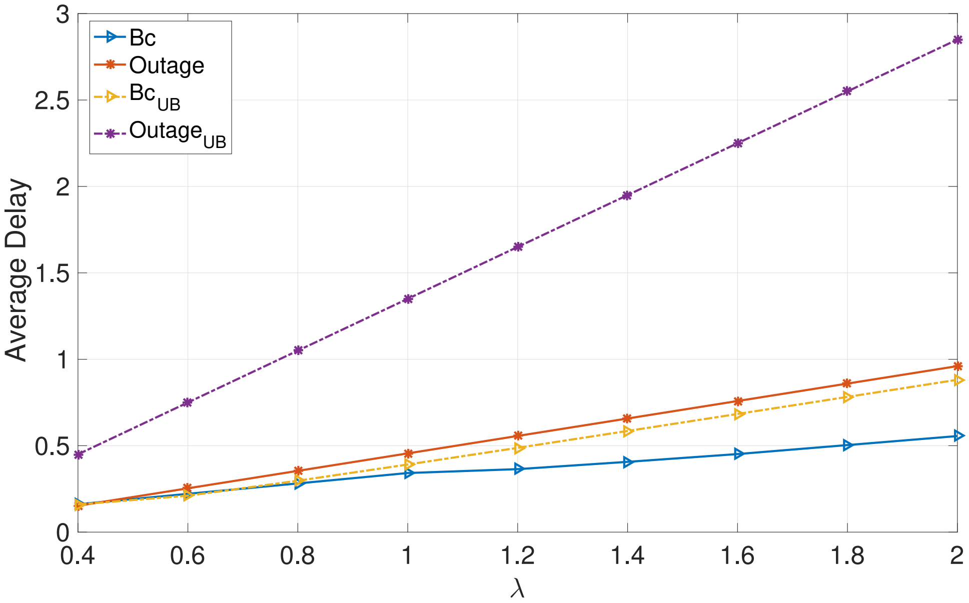

Figure 2. Deterministic: Asymmetric.

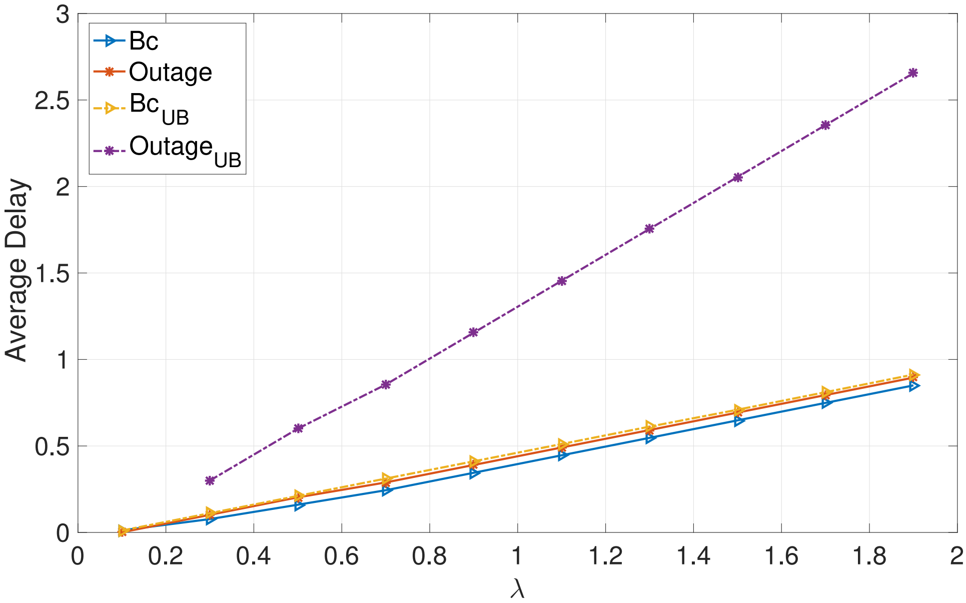

Figure 3. Poisson: Symmetric.

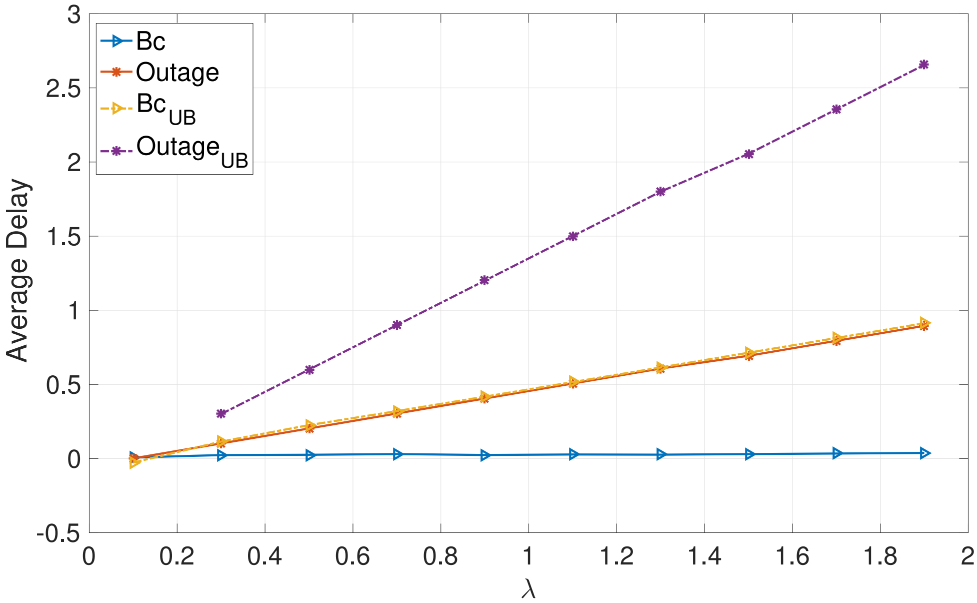

Figure 4. Poisson: Asymmetric.

The numerical evaluations support the analysis, demonstrating that the proposed broadcast approach significantly enhances the average delays of both users in the moderate and high SNR regimes for moderate and high arrival rates.

7. Concluding Remarks

In this paper, a non-orthogonal multi-access broadcast approach is employed, in which each user splits its information stream into a finite number of encoded layers, each adapted to one possible network state, serving as an outage-free low-latency transmission scheme. In particular, the average queuing delay of each user under the proposed multi-access approach is analyzed for different arrival processes at each transmitter. First, for deterministic arrivals, closed-form lower and upper bounds on the average delay are derived analytically. Secondly, for Poisson arrival rates, the average queuing delay is characterized in a closed form. The latency advantage of the proposed approach compared to the single-layer transmission is shown analytically. Finally, we note that in this paper, our focus has been on the discrete channel models since it provides a setting based on which the key ideas (specifically information layering and decoding strategy) can be described clearly and in detail. In order to gain insight into the behavior in the continuous channel models, by increasing the number of channel states in the limit of an infinite number of states, the models converge to a continuous model, and the codebook assignments and decoding strategy converge to their counterparts for continuous channels (larger number of codebooks with low rates).

Author Contributions: All authors have contributed equally to the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: The work of A. Tajer has been supported in part by the U.S. Nationa Science Foundation grant ECCS-1933107. The work of S. Shamai (Shitz) has been supported by the European Union's Horizon 2020 Research And Innovation Programme, grant agreement no. 694630.

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

Appendix A. Constants of Theorem 1

$\forall i \in \mathcal{N}$:

$$d^i_{10}(\phi) \triangleq C\left(\alpha_1\beta^i_{10}, N\alpha_1 - \sum_{j=1}^{i}\alpha_1\beta^j_{10}\right) . \tag{A1}$$

$\forall m \in \mathcal{S}_k$:

$$d^m_{10}(\mathcal{S}_k) \triangleq C\left(\alpha_2\beta^m_{10}, (N-k)\alpha_1 + k\alpha_2 - \sum_{j\in\mathcal{S}_k, j\le k(m)} \alpha_2(1-\beta^j_{10})\right) , \tag{A2}$$

$$d^m_{11}(\mathcal{S}_k) \triangleq C\left(\alpha_2\beta^m_{11}, (N-k)\alpha_1 + k\alpha_2 - \sum_{j\in\mathcal{S}_k}\alpha_2\beta^j_{10} - \sum_{j\notin\mathcal{S}_k}\alpha_1\beta^j_{1k} - \sum_{j\in\mathcal{S}_k, j\le k(m)}\alpha_2\beta^j_{11})\right) , \tag{A3}$$

$$\begin{aligned} d^m_{21}(\mathcal{S}_k) \triangleq C\Bigg(&\alpha_2\beta^m_{21}, (N-k)\alpha_1 + k\alpha_2 - \sum_{j\in\mathcal{S}_k}\alpha_2(\beta^j_{10}+\beta^j_{11}) \\ &- \sum_{j\notin\mathcal{S}_k}\alpha_1(\beta^j_{1k}+\beta^j_{2(k-1)}) - \sum_{j\in\mathcal{S}_k, j\le k(m)}\alpha_2\beta^j_{21})\Bigg) \end{aligned} \tag{A4}$$

$\forall m \in \mathcal{S}_k$ and $\ell \in \{1,\ldots,k\}$:

$$\begin{aligned} d^m_{1\ell}(\mathcal{S}_k) \triangleq C\Bigg(&\alpha_2\beta^m_{1\ell}, (N-k)\alpha_1 + k\alpha_2 - \sum_{j\in\mathcal{S}_k}\alpha_2\beta^j_{10} - \sum_{j\in\mathcal{S}_k}\sum_{i=1}^{\ell-1}\alpha_2(\beta^j_{1i}+\beta^j_{2i}) \\ &- \sum_{j\notin\mathcal{S}_k}\alpha_1\sum_{i=1}^{\ell-1}(\beta^j_{1(k-i+1)}+\beta^j_{2(k-i+1)}) - \sum_{j\in\mathcal{S}_k, j\le k(m)}\alpha_2\beta^j_{1\ell}\Bigg) , \end{aligned} \tag{A5}$$

$$\begin{aligned} d^m_{2\ell}(\mathcal{S}_k) \triangleq C\Bigg(&\alpha_2\beta^m_{2\ell}, (N-k)\alpha_1 + k\alpha_2 - \sum_{j\in\mathcal{S}_k}\alpha_2\beta^j_{10} - \sum_{j\in\mathcal{S}_k}\sum_{i=1}^{\ell}\alpha_2\beta^j_{1i} - \sum_{j\in\mathcal{S}_k}\sum_{i=1}^{\ell-1}\alpha_2\beta^j_{2i} \\ &- \sum_{j\notin\mathcal{S}_k}\alpha_1\sum_{i=1}^{\ell}(\beta^j_{1(k-i+1)}+\beta^j_{2(k-i+1)}) - \sum_{j\in\mathcal{S}_k, j\le k(m)}\alpha_2\beta^j_{2\ell}\Bigg) . \end{aligned} \tag{A6}$$

$\forall n \notin \mathcal{S}_k$ and $\ell \in \{1, \ldots, k\}$:

$$\begin{aligned} d^n_{1(k-\ell+1)}(\mathcal{S}_k) \triangleq C\Bigg(& \alpha_1\beta^n_{1(k-\ell+1)}, (N-k)\alpha_1 + k\alpha_2 - \sum_{j\in\mathcal{S}_k} \alpha_2\beta^j_{10} - \sum_{j\in\mathcal{S}_k}\sum_{i=1}^{\ell-1} \alpha_2(\beta^j_{1i} + \beta^j_{2i}) \\ & - \sum_{j\notin\mathcal{S}_k} \alpha_1 \sum_{i=1}^{\ell-1}(\beta^j_{1(k-i+1)} + \beta^j_{2(k-i+1)}) - \sum_{j\notin\mathcal{S}_k, j\leq\bar{k}(n)} \alpha_1\beta^j_{1(k-\ell+1)})\Bigg)\,, \end{aligned} \tag{A7}$$

$$\begin{aligned} d^n_{2(k-\ell)}(\mathcal{S}_k) \triangleq C\Bigg(& \alpha_1\beta^n_{2(k-\ell)}, (N-k)\alpha_1 + k\alpha_2 - \sum_{j\in\mathcal{S}_k} \alpha_2\beta^j_{10} - \sum_{j\in\mathcal{S}_k}\sum_{i=1}^{\ell} \alpha_2\beta^j_{1i} - \sum_{j\in\mathcal{S}_k}\sum_{i=1}^{\ell-1} \alpha_2\beta^j_{2i} \\ & - \sum_{j\notin\mathcal{S}_k}\sum_{i=1}^{\ell} \alpha_1\beta^j_{1(k-i+1} - \sum_{j\notin\mathcal{S}_k}\sum_{i=1}^{\ell-1} \alpha_1\beta^j_{2(k-i+1)} - \sum_{j\notin\mathcal{S}_k, j\leq\bar{k}(n)} \alpha_1\beta^j_{2(k-\ell)})\Bigg)\,. \end{aligned} \tag{A8}$$

Appendix B. Proof of Theorem 1

The rate region characterized in Theorem 1 is achievable by employing the layering scheme in Section 3.1 at each transmitter combined with the successive decoding strategy in Algorithm 1. Recall that the maximum rate of codeword U^i_{jk}, for each user $i \in \mathcal{N}$ channel $j \in \{1, 2\}$, and $k \in \{0 \cup \mathcal{N}\}$, is bounded by the minimum achievable rate for that codebook in all combined network states during which it is decoded.

We define $\mathcal{S}$ as the set of users' indices that are experiencing a *strong* states. This set is known to the receiver. Accordingly, we define $\mathcal{S}_k$ as a realization of S that contains exactly k users, i.e., k users have strong channels and $N-k$ users have weak channels. Next, we discuss $\mathcal{S}_0$ and $\mathcal{S}_k$ for $k \in \mathcal{N}$, separately.

$|\mathcal{S}| = 0$**:** All channels are weak

In the event of a network state with all channels in the weak state, $h_i^2 = \alpha_1$, $\forall i \in \mathcal{N}$, the receiver decodes only one layer per user. Specifically, it decodes $\{U^i_{10} : i \in \mathcal{N}\}$. It performs successive decoding, starting from user 1 and continuing in the ascending order of users' indices. In order to successfully decode layers $\{U^i_{10} : i \in \mathcal{N}\}$, the rate of each layer $i \in \mathcal{N}$ should satisfy:

$$\forall i \in \mathcal{N}: \qquad R^i_{10} \leq C\left(\alpha_1\beta^i_{10}, N\alpha_1 - \sum_{j=1}^{i} \alpha_1\beta^j_{10}\right) \triangleq d^i_{10}(\phi). \tag{A9}$$

Note that the second argument $C(x, y)$ represents the undecoded layers that will be treated as interference for layer U^i_{10}. Hence, based on the successive decoding procedure, when the receiver decodes U^i_{10}, layers U^i_{10} for users $j \in \{1, \ldots, i-1\}$ have already been decoded. Thus, their interference is subtracted from the total transmitted signal, accounted by the term $1 - \beta^j_{10}$. On the other hand, none of the layers transmitted by users $j \in \{i+1, \ldots, N\}$ have been decoded yet, which is accounted by the term $(N-i)\alpha_1$.

Next, we will characterize upper bounds on the achievable rates of all the layers decoded when there are exactly k users with strong channels, i.e., $|\mathcal{S}| = k$.

$|\mathcal{S}| = k$**:** k channels are strong

As discussed earlier, when $|\mathcal{S}| = k$, the receiver employs $4k+1$ decoding stages. For this purpose, the set of codebooks $\{U^i_{j\ell} : i \in \mathcal{N}\}$ is partitioned to two sets

$$\mathcal{P}_{j\ell} \triangleq \{U^i_{j\ell} : i \in \mathcal{S}\} \qquad \text{and} \qquad \mathcal{Q}_{j\ell} \triangleq \{U^i_{j\ell} : i \notin \mathcal{S}\}\,, \tag{A10}$$

rendering a total of $4k+1$ partitions for different $j \in \{1, 2\}$ and $\ell \in \{0, \ldots, k\}$. The decoding strategy, decodes one message from each of these, except for the partition $\{U^i_{2k} : i \notin \mathcal{S}\}$. The decoding strategy works as follows. We create the following two sequences of sets:

$$\mathcal{P} \triangleq \{\mathcal{P}_{10}, \mathcal{P}_{11}, \mathcal{P}_{21}, \ldots, \mathcal{P}_{2(k-1)}, \mathcal{P}_{1k}, \}\,, \tag{A11}$$

$$\mathcal{Q} \triangleq \{\mathcal{Q}_{1k}, \mathcal{Q}_{2(k-1)}, \mathcal{Q}_{1(k-1)}, \ldots, \mathcal{Q}_{11}, \mathcal{Q}_{10}, \}\,. \tag{A12}$$

The decoding strategy selects codebooks by alternating between $\mathcal{P}$ and $\mathcal{Q}$ an an ascending order and decodes exactly one codebook from each. This results in $4k$ coding stages. Finally, the codebooks in $\{U^i_{2k} : i \in \mathcal{S}\}$ are decoded as the last stage, i.e., stage $4k + 1$.

- **Decoding stage** 1**:**
  We start by decoding the layers $\mathcal{P}_{10} \triangleq \{U^i_{10} : i \in \mathcal{S}\}$. Recall that $\mathcal{S}_k$ was defined as an ordered set of these users. The codebooks will be decoded sequentially in this order. When $m \in \mathcal{S}_k$, we denote the position of m in $\mathcal{S}_k$ by $k(m)$. Hence, $\forall m \in \mathcal{S}_k$

$$R^m_{10} \leq C\left(\alpha_2\beta^m_{10}, (N-k)\alpha_1 + k\alpha_2 - \sum_{j\in\mathcal{S}_k, j\leq k(m)} \alpha_2(1-\beta^j_{10})\right) \triangleq d^m_{10}(\mathcal{S}_k)\,. \tag{A13}$$

- **Decoding stage** 2**:**
  Next, we sequentially decode the layers in $\mathcal{Q}_{1k} = \{U^i_{1k} : i \notin \mathcal{S}\}$, which involves layers U^i_{1k} of users with weak channels. When $n \notin \mathcal{S}_k$, we denote the position of n in the ordered set $\mathcal{N}\backslash\mathcal{S}_k$ by $\bar{k}(n)$. Hence, $\forall n \notin \mathcal{S}_k$

$$R^n_{1k} \leq C\left(\alpha_1\beta^n_{1k}, (N-k)\alpha_1 + k\alpha_2 - \sum_{j\in\mathcal{S}_k} \alpha_2\beta^j_{10} - \sum_{j\notin\mathcal{S}_k, j<\bar{k}(n)} \alpha_1\beta^j_{1k}\right) \triangleq d^n_{1k}(\mathcal{S}_k)\,. \tag{A14}$$

- **Decoding stage** 3**:**
  In the third stage, the codebooks in $\mathcal{P}_{10}$ and $\mathcal{Q}_{1k}$ are already decoded. We continue by sequentially decoding the set of codebooks in $\mathcal{P}_{11} \triangleq \{U^i_{11} : i \in \mathcal{S}\}$. Hence, $\forall m \in \mathcal{S}_k$

$$R^m_{11} \leq C\left(\alpha_2\beta^m_{11}, (N-k)\alpha_1 + k\alpha_2 - \sum_{j\in\mathcal{S}_k} \alpha_2\beta^j_{10} - \sum_{j\notin\mathcal{S}_k} \alpha_1\beta^j_{1k} - \sum_{j\in\mathcal{S}_k, j\leq k(m)} \alpha_2\beta^j_{11})\right) \triangleq d^m_{11}(\mathcal{S}_k)\,. \tag{A15}$$

- **Decoding stage** 4**:**
  The decoding process continues by sequentially decoding the codebooks in $\mathcal{Q}_{2(k-1)} = \{U^i_{2(k-1)} : i \notin \mathcal{S}\}$, while the codebooks of $\mathcal{P}_{10}$, $\mathcal{Q}_{1k}$, and $\mathcal{P}_{11}$ are already decoded. Hence, for $n \notin \mathcal{S}_k$

$$\begin{aligned} R^n_{2(k-1)} \leq C\Bigg(&\alpha_1\beta^n_{2(k-1)}, (N-k)\alpha_1 + k\alpha_2 - \sum_{j\in\mathcal{S}_k} \alpha_2(\beta^j_{10} + \beta^j_{11}) \\ &- \sum_{j\notin\mathcal{S}_k} \alpha_1\beta^j_{1k} - \sum_{j\notin\mathcal{S}_k, j\leq\bar{k}(n)} \alpha_1\beta^j_{2(k-1)})\Bigg) \triangleq d^n_{2(k-1)}(\mathcal{S}_k)\,. \end{aligned} \tag{A16}$$

- **Decoding stage** 5**:**
  This stage sequentially decodes the codebooks $\mathcal{P}_{21}$. For all $m \in \mathcal{S}^i_k$ we have

$$\begin{aligned} R^m_{21} \leq C\Bigg(&\alpha_2\beta^m_{21}, (N-k)\alpha_1 + k\alpha_2 - \sum_{j\in\mathcal{S}_k} \alpha_2(\beta^j_{10} + \beta^j_{11}) \\ &- \sum_{j\notin\mathcal{S}_k} \alpha_1(\beta^j_{1k} + \beta^j_{2(k-1)}) - \sum_{j\in\mathcal{S}_k, j\leq k(m)} \alpha_2\beta^j_{21})\Bigg) \triangleq d^m_{21}(\mathcal{S}_k)\,. \end{aligned} \tag{A17}$$

- **Decoding stage** 6**:**
  This stage sequentially decodes the codebooks in $\mathcal{Q}_{1(k-1)}$. Hence, $\forall n \notin \mathcal{S}_k$ we have

$$R^n_{1(k-1)} \leq C\Bigg(\alpha_1\beta^n_{1(k-1)}, (N-k)\alpha_1 + k\alpha_2 - \sum_{j\in\mathcal{S}_k}\alpha_2(\beta^j_{10}+\beta^j_{11}+\beta^j_{21}) - \sum_{j\notin\mathcal{S}_k}\alpha_1(\beta^j_{1k}+\beta^j_{2(k-1)}) - \sum_{j\notin\mathcal{S}_k, j\leq\bar{k}(n)}\alpha_1\beta^j_{1(k-1)})\Bigg) \triangleq d^n_{1(k-1)}(\mathcal{S}_k)\ . \tag{A18}$$

- **Decoding stages $\{2,\ldots,4k+1\}$:**
  Following the pattern of the previous decoding stages, in general in the stage $\{2,\ldots,4k\}$, we decode the codebooks according to the following schedule, for $\ell\in\{1,\ldots,k\}$:

$$\begin{array}{lll} \text{codebooks in } \mathcal{Q}_{1(k-\ell+1)} & \text{stage} & 4\ell-2 \\ \text{codebooks in } \mathcal{P}_{1\ell} & \text{stage} & 4\ell-1 \\ \text{codebooks in } \mathcal{Q}_{2(k-\ell)} & \text{stage} & 4\ell \\ \text{codebooks in } \mathcal{P}_{2\ell} & \text{stage} & 4\ell+1 \end{array} \tag{A19}$$

  Accordingly, we obtain the following rate constraints.

- **Decoding stage $4\ell-2$:**

  By sequentially decoding the messages in $\mathcal{Q}_{1(k-\ell+1)}$, $\forall n\notin\mathcal{S}_k$ we have

$$R^n_{1(k-\ell+1)} \leq C\Bigg(\alpha_1\beta^n_{1(k-\ell+1)}, (N-k)\alpha_1 + k\alpha_2 - \sum_{j\in\mathcal{S}_k}\alpha_2\beta^j_{10} - \sum_{j\in\mathcal{S}_k}\sum_{i=1}^{\ell-1}\alpha_2(\beta^j_{1i}+\beta^j_{2i}) - \sum_{j\notin\mathcal{S}_k}\alpha_1\sum_{i=1}^{\ell-1}(\beta^j_{1(k-i+1)}+\beta^j_{2(k-i+1)}) - \sum_{j\notin\mathcal{S}_k, j\leq\bar{k}(n)}\alpha_1\beta^j_{1(k-\ell+1)})\Bigg) \triangleq d^n_{1(k-\ell+1)}(\mathcal{S}_k)\ . \tag{A20}$$

- **Decoding stage $4\ell-1$:**
  By sequentially decoding the messages in $\mathcal{P}_{1\ell}$, $\forall m\in\mathcal{S}_k$ we have

$$R^m_{1\ell} \leq C\Bigg(\alpha_2\beta^m_{1\ell}, (N-k)\alpha_1 + k\alpha_2 - \sum_{j\in\mathcal{S}_k}\alpha_2\beta^j_{10} - \sum_{j\in\mathcal{S}_k}\sum_{i=1}^{\ell-1}\alpha_2(\beta^j_{1i}+\beta^j_{2i}) - \sum_{j\notin\mathcal{S}_k}\alpha_1\sum_{i=1}^{\ell-1}(\beta^j_{1(k-i+1)}+\beta^j_{2(k-i+1)}) - \sum_{j\in\mathcal{S}_k, j\leq k(m)}\alpha_2\beta^j_{1\ell}\Bigg) \triangleq d^m_{1\ell}(\mathcal{S}_k)\ . \tag{A21}$$

- **Decoding stage 4ℓ:**

  By sequentially decoding the messages in $\mathcal{Q}_{1(k-\ell)}$, $\forall n\notin\mathcal{S}_k$ we have

$$R^n_{2(k-\ell)} \leq C\Bigg(\alpha_1\beta^n_{2(k-\ell)}, (N-k)\alpha_1 + k\alpha_2 - \sum_{j\in\mathcal{S}_k}\alpha_2\beta^j_{10} - \sum_{j\in\mathcal{S}_k}\sum_{i=1}^{\ell}\alpha_2\beta^j_{1i} - \sum_{j\in\mathcal{S}_k}\sum_{i=1}^{\ell-1}\alpha_2\beta^j_{2i} - \sum_{j\notin\mathcal{S}_k}\sum_{i=1}^{\ell}\alpha_1\beta^j_{1(k-i+1} - \sum_{j\notin\mathcal{S}_k}\sum_{i=1}^{\ell-1}\alpha_1\beta^j_{2(k-i+1)} - \sum_{j\notin\mathcal{S}_k, j\leq\bar{k}(n)}\alpha_1\beta^j_{2(k-\ell)})\Bigg) \triangleq d^n_{2(k-\ell)}(\mathcal{S}_k)\ . \tag{A22}$$

- **Decoding stage $4\ell+1$:**
  By sequentially decoding the messages in $\mathcal{P}_{2\ell}$, $\forall m\in\mathcal{S}_k$ we have

$$\begin{aligned}R^m_{2\ell} \leq C\Bigg(&\alpha_2\beta^m_{2\ell}, (N-k)\alpha_1 + k\alpha_2 - \sum_{j\in\mathcal{S}_k} \alpha_2\beta^j_{10} - \sum_{j\in\mathcal{S}_k}\sum_{i=1}^{\ell} \alpha_2\beta^j_{1i} - \sum_{j\in\mathcal{S}_k}\sum_{i=1}^{\ell-1} \alpha_2\beta^j_{2i} \\ &- \sum_{j\notin\mathcal{S}_k} \alpha_1 \sum_{i=1}^{\ell}(\beta^j_{1(k-i+1)} + \beta^j_{2(k-i+1)}) - \sum_{j\in\mathcal{S}_k, j\leq k(m)} \alpha_2\beta^j_{2\ell}\Bigg) \triangleq d^m_{2\ell}(\mathcal{S}_k)\,. \end{aligned} \tag{A23}$$

Given the upper bounds on the individual achievable rates of $U^i_{jk}, \forall i \in \mathcal{N}, j \in 1,2$, $k \in \{0 \cup \mathcal{N}\}$, the maximum achievable rate of U^i_{jk} is bounded my the minimum upper bound among all the network states within which it is decoded.

Appendix C. Proof of Theorem 2

By applying a change of variable to each term and taking the integral $\int_0^\infty e^{-sq} dF_1(q)$ as a common factor, $L_1(s)$ can be expressed as

$$L_1(s) = \frac{F_1(0) - \int_{0^+}^{(\sum_{ij} R^1_{ij} - \lambda_1)} e^{-s(q+(\lambda_1 - \sum_{ij} R^1_{ij}))} dF_1(q)}{1 - [\bar{p}_1\bar{p}_2 e^{-s(\lambda_1 - \sum_{ij} R^1_{ij})} + \bar{p}_1 p_2 e^{-s(\lambda_1 - R^1_{11} - R^1_{21}))} + p_1\bar{p}_2 e^{-s(\lambda_1 - R^1_{11} - R^1_{12}))} + p_1 p_2 e^{-s(\lambda_1 - R^1_{11}))}]} \tag{A24}$$

Further, by using the definition of $F_1(0) = \bar{p}_1\bar{p}_2 F_1(q - (\lambda_1 - \sum_{ij} R^1_{ij}))$ and multiplying the numerator and denominator of (A24) by a common factor, $e^{-s(\sum_{ij} R^1_{ij} - \lambda_1)}$, we have.

$$\begin{aligned} L_1(s) &= \frac{\bar{p}_1\bar{p}_2[\int_0^{(\sum_{ij} R^1_{ij} - \lambda_1)} e^{-s(\sum_{ij} R^1_{ij} - \lambda_1)} - e^{-sq} dF_1(q)]}{e^{-s(\sum_{ij} R^1_{ij} - \lambda_1)} - [\bar{p}_1\bar{p}_2 + \bar{p}_1 p_2 e^{-s(R^1_{12} + R^1_{22}))} + p_1\bar{p}_2 e^{-s(R^1_{21} + R^1_{22}))} + p_1 p_2 e^{-s(R^1_{21} + R^1_{12} + R^1_{22}))}]} \\ &\triangleq \frac{D_1(s)}{N_1(s)}\,. \end{aligned} \tag{A25}$$

It can be readily noticed from (A25) that $\lim_{s\to 0} D_1(s) = \lim_{s\to 0} N_1(s) = 0$, therefore we apply L'hopital's limit rule on (A25) to arrive at

$$\mathbb{E}[Q_1] = \lim_{s\to 0} \frac{D^{''}_{Q_1}(s) - N^{''}_{Q_1}(s)}{2D^{'}_{Q_1}(s)}\,. \tag{A26}$$

Finally, we evaluate the terms $D^{''}_{Q_1}(s)$, $N^{''}_{Q_1}(s)$ and $D^{'}_{Q_1}(s)$ where we have

$$\lim_{s\to 0} D^{'}_{Q_1}(s) = -(\sum_{ij} R^1_{ij} - \lambda_1) + \bar{p}_1 p_2(R^1_{12} + R^1_{22}) + p_1\bar{p}_2(R^1_{12} + R^1_{22}) + p_1 p_2(R^1_{12} + R^1_{21} + R^1_{22}) \tag{A27}$$

$$\lim_{s\to 0} D^{''}_{Q_1}(s) = (\sum_{ij} R^1_{ij} - \lambda_1)^2 - \bar{p}_1 p_2(R^1_{12} + R^1_{22})^2 - p_1\bar{p}_2(R^1_{12} + R^1_{22})^2 - p_1 p_2(R^1_{12} + R^1_{21} + R^1_{22})^2\,, \tag{A28}$$

and

$$\lim_{s\to 0} N^{''}_{Q_1}(s) = \bar{p}_1\bar{p}_2 \int_0^{(\sum_{ij} R^1_{ij} - \lambda_1)} [(\sum_{ij} R^1_{ij} - \lambda_1)^2 - q^2] dF_1(q)\,. \tag{A29}$$

Finally, by using $\lim_{s\to 0} D^{'}_1(s) = \lim_{s\to 0} N^{'}_1(s)$, the second derivative of the numerator term can be upper bounded by replacing $(\sum_{ij} R^1_{ij} - \lambda_1 + q)$ by $2(\sum_{ij} R^1_{ij} - \lambda_1)$ arriving at

$$\lim_{s\to 0} N''_{Q_1}(s) \leq 2(\sum_{ij} R^1_{ij} - \lambda_1)\Bigg(\sum_{ij} R^1_{ij} - \lambda_1 \\ -\bar{p}_1 p_2 (R^1_{12} + R^1_{22}) - p_1 \bar{p}_2 (R^1_{12} + R^1_{22}) - p_1 p_2 (R^1_{12} + R^1_{21} + R^1_{22})\Bigg)\,. \tag{A30}$$

Next, we leverage (A26) reaching

$$\begin{aligned} \mathbb{E}[Q_i] &\geq \frac{1}{2} \textstyle\sum_{j=1}^2 \sum_{k=0}^N R^i_{jk} - \frac{\lambda_i}{2} - \frac{N_i}{D_i}\,, \\ \mathbb{E}[Q_i] &\leq \textstyle\sum_{j=1}^2 \sum_{k=0}^N R^i_{jk} - \lambda_i - \frac{N_i}{D_i}\,, \end{aligned} \tag{A31}$$

where

$$\begin{aligned} N_i &\triangleq -\left(\textstyle\sum_{j=1}^2 \sum_{k=0}^N R^i_{jk} - \lambda_i\right)^2 \\ &+ \mathbb{P}\left[\mathcal{E}\left(\bar{\mathcal{S}}^i_0\right)\right]\left(\textstyle\sum_{j=1}^2 \sum_{k=1}^N R^i_{jk}\right)^2 \\ &+ \textstyle\sum_{\ell=1}^{N-1} \mathbb{P}[\mathcal{E}(\mathcal{S}^i_\ell)] \cdot \left(\sum_{j=1}^2 \sum_{k=\ell+1}^N R^i_{jk} + R^i_{1\ell}\right)^2 \\ &+ \textstyle\sum_{\ell=1}^{N-1} \mathbb{P}\left[\mathcal{E}\left(\bar{\mathcal{S}}^i{}_\ell\right)\right] \cdot \left(\sum_{j=1}^2 \sum_{k=\ell+1}^N R^i_{jk} + R^i_{2\ell}\right)^2 \\ D_i &\triangleq 2(\mathbb{E}[r_i] - \lambda_i)\,. \end{aligned} \tag{A32}$$

Appendix D. Proof of Theorem 3

In this Appendix, we base the proof of Theorem 3 on two main steps. First, we characterize a lower bound on the average achievable rate of each user i using a single layer per user (outage approach). Secondly, we derive the rate of increase of the average achievable delay with respect to the average arrival rate λ_i (first-order derivative) for the delay upper bound of the multi-layer approach to that of the delay lower bound of the outage approach. Finally, under a fixed average achievable rate among both approaches, we show that the proposed approach outperforms the single layer outage approach.

Recalling the recursive expression for Q_i in terms of the variable Z_i in (3), a recursive form of $F_i(q)$ can be expressed as follows [52,53]

$$F_i(q) = \begin{cases} 0, & q < 0 \\ \int_{-\infty}^{q} F_i(q-\tau) dF_{Z_i}(\tau), & q \geq 0\,, \end{cases} \tag{A33}$$

where $dF_{Z_i}(z)$ denote pdf of Z_i.

At the end of every transmission block, the change in queue size i, Z_i, is primarily determined by the difference between the data arrival λ_i and the fixed rate successfully decoded at the receiver, which in turn is determined by the combined network state. Consequently, $dF_{Z_i}(z)$ can be expressed by

$$dF_{Z_i}(z) = P_{\text{out}} \delta(z - \lambda_i + R_{\text{F}})\,. \tag{A34}$$

We remark that in order to guarantee the stability of every queue i, we assume that the arrival rate λ_i is less that the average achievable rate (service rate of the queue), i.e.,

$$\lambda_i < P_{\text{out}} R_{\text{F}},\ \forall i \in \mathcal{N}\,. \tag{A35}$$

Combining (12) and (13), an explicit expression for $F_i(q), \forall i \in \mathcal{N}$ is given by

$$F_i(q) = \\ \begin{cases} 0\,, & \forall q < 0 \\ P_{\text{out}} F_i(q - \lambda_i + R_{\text{F}})\,, & \forall q \geq 0 \end{cases}. \tag{A36}$$

Finally, we evaluate the terms $D''_{\mathrm{Q}_1}(s)$ and $N''_{\mathrm{Q}_1}(s)$ where we have

$$\lim_{s\to 0} D''_{\mathrm{Q}_1}(s) = (R_{\mathrm{F}} - \lambda_i)^2 - (1 - P_{\mathrm{out}})R_{\mathrm{F}}^2, \tag{A37}$$

and

$$\lim_{s\to 0} N''_{\mathrm{Q}_1}(s) = P_{\mathrm{out}} \int_0^{(R_{\mathrm{F}}-\lambda_1)} [(R_{\mathrm{F}} - \lambda_1)^2 - q^2] dF_1(q)\,. \tag{A38}$$

Finally, by using $\lim_{s\to 0} D'_1(s) = \lim_{s\to 0} N'_1(s)$, the second derivative of the numerator term can be lower bounded by replacing $(F_{\mathrm{F}} - \lambda_1 + q)$ by $(R_{\mathrm{F}} - \lambda_1)$ arriving at

$$\lim_{s\to 0} N''_{\mathrm{Q}_1}(s) \geq (R_{\mathrm{F}} - \lambda_1)(R_{\mathrm{F}} - \lambda_1 - P_{\mathrm{out}}R_{\mathrm{F}})\,. \tag{A39}$$

and substitute (A26) reaching

$$\mathbb{E}[Q_i] \geq \frac{1}{2}R_{\mathrm{F}} - \frac{\lambda_i}{2} - \frac{N_i}{D_i}, \tag{A40}$$

where

$$\begin{aligned} N_i &\triangleq -(R_{\mathrm{F}} - \lambda_i)^2 + (1 - P_{\mathrm{out}})R_{\mathrm{F}}^2 \\ D_i &\triangleq P_{\mathrm{out}}R_{\mathrm{F}} - \lambda_i\,. \end{aligned} \tag{A41}$$

By taking the derivative of the upper/lower bounds derived above we reach

$$\frac{\partial U_B}{\partial \lambda_i} = -1 - \frac{\sum_{j=1}^{2}\sum_{k=0}^{N} R_{jk}^i - \lambda_i}{\mathbb{E}[r_i] - \lambda_i} - 2\frac{N_i}{D_i^2}, \tag{A42}$$

$$\frac{\partial L_B}{\partial \lambda_i} = -1 - \frac{R_{\mathrm{F}} - \lambda_i}{P_{\mathrm{out}}R_{\mathrm{F}} - \lambda_i} - 2\frac{-(R_{\mathrm{F}} - \lambda_i)^2 + (1 - P_{\mathrm{out}})R_{\mathrm{F}}^2}{P_{\mathrm{out}}R_{\mathrm{F}} - \lambda_i}\,. \tag{A43}$$

References

1. *Study on Scenarios and Requirements for Next Generation Access Technologies*; ETSI: Sophia Antipolis, France, 2017.
2. Ding, Z.; Liu, Y.; Choi, J.; Sun, Q.; Elkashlan, M.; Chih-Lin, I.; Poor, H.V. Application of non-orthogonal multiple access in LTE and 5G networks. *IEEE Commun. Mag.* **2017**, *55*, 185–191. [CrossRef]
3. Ding, Z.; Lei, X.; Karagiannidis, G.K.; Schober, R.; Yuan, J.; Bhargava, V.K. A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends. *IEEE J. Sel. Areas Commun.* **2017**, *35*, 2181–2195. [CrossRef]
4. Ding, Z.; Yang, Z.; Fan, P.; Poor, H.V. On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users. *IEEE Signal Process. Lett.* **2014**, *21*, 1501–1505. [CrossRef]
5. Islam, S.R.; Avazov, N.; Dobre, O.A.; Kwak, K.-S. Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges. *IEEE Commun. Surv. Tutor.* **2016**, *19*, 721–742. [CrossRef]
6. Benjebbovu, A.; Li, A.; Saito, Y.; Kishiyama, Y.; Harada, A.; Nakamura, T. System-level performance of downlink NOMA for future LTE enhancements. In Proceedings of the IEEE Global Communications Conference Workshops, Atlanta, GA, USA, 9–13 December 2013; pp. 66–70.
7. Wang, W.; Liu, Y.; Luo, Z.; Jiang, T.; Zhang, Q.; Nallanathan, A. Toward cross-layer design for non-orthogonal multiple access: A quality-of-experience perspective. *IEEE Wirel. Commun.* **2018**, *25*, 118–124. [CrossRef]
8. Condoluci, M.; Dohler, M.; Araniti, G.; Molinaro, A.; Sachs, J. Enhanced radio access and data transmission procedures facilitating industry-compliant machine-type communications over LTE-based 5G networks. *IEEE Wirel. Commun.* **2016**, *23*, 56–63. [CrossRef]
9. Bennis, M.; Debbah, M.; Poor, H.V. Ultrareliable and low-latency wireless communication: Tail, risk, and scale. *Proc. IEEE* **2018**, *106*, 1834–1853. [CrossRef]
10. Shamai, S.; Steiner, A. A broadcast approach for a single-user slowly fading MIMO channel. *IEEE Trans. Inf. Theory* **2003**, *49*, 2617–2635. [CrossRef]
11. Shamai, S. A broadcast strategy for the Gaussian slowly fading channel. In Proceedings of the IEEE International Symposium Information Theory, Ulm, Germany, 29 June–4 July 1997; p. 150.
12. Tajer, A.; Steiner, A.; Shamai, S. The broadcast approach in communication networks. *Entropy* **2021**, *23*, 120. [CrossRef]
13. Zohdy, M.; Tajer, A. Broadcast Approach for the Single-user Energy Harvesting Channel. *IEEE Trans. Commun.* **2019**, *67*, 3192–3204. [CrossRef]

14. Shamai, S. A broadcast approach for the multiple-access slow fading channel. In Proceedings of the IEEE International Symposium Information Theory, Sorrento, Italy, 25–30 June 2000; p. 128.
15. Minero, P.; Tse, D.N.C. A broadcast approach to multiple access with random states. In Proceedings of the IEEE International Symposium Information Theory, Nice, France, 24–29 June 2007; pp. 2566–2570.
16. Minero, P.; David, N.; Franceschetti, M. A broadcast approach to random access. In Proceedings of the IEEE Information Theory Workshop, Taormina, Italy, 11–16 October 2009; pp. 615–619.
17. Kazemi, S.; Tajer, A. Multiaccess communication via a broadcast approach adapted to the multiuser channel. *IEEE Trans. Commun.* **2018**, *66*, 3341–3353. [CrossRef]
18. Zohdy, M.; Tajer, A.; Shamai, S. Broadcast Approach to Multiple Access with Local CSIT. *IEEE Trans. Commun.* **2019**, *67*, 7483–7498. [CrossRef]
19. Kazemi, S.; Tajer, A. A broadcast approach to multiple access adapted to the multiuser channel. In Proceedings of the IEEE International Symposium on Information Theory, Aachen, Germany, 25–30 June 2017.
20. Zohdy, M.; Kazemi, S.; Tajer, A. A broadcast approach to multiple access with partial CSIT. In Proceedings of the IEEE Global Communications Conference, Abu Dhabi, United Arab Emirates, 9–13 December 2018.
21. Zohdy, M.; Tajer, A.; Shamai, S. Interference Management without CSIT: A Broadcast Approach. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Los Angeles, CA, USA, 21–26 June 2020.
22. Zohdy, M.; Tajer, A.; Shamai, S. Distributed Interference Management: A Broadcast Approach. *IEEE Trans. Commun.* **2021**, *69*, 149–163. [CrossRef]
23. Ye, N.; Wang, A.; Li, X.; Liu, W.; Hou, X.; Yu, H. Rate-adaptive multiple access for uplink grant-free transmission. *Wirel. Commun. Mob. Comput.* **2018**, *2018*, 1–21. [CrossRef]
24. Steiner, A.; Shamai, S. On queueing and multilayer coding. *IEEE Trans. Inf. Theory* **2010**, *56*, 2392–2415. [CrossRef]
25. Clerckx, B.; Mao, Y.; Jorswieck, E.A.; Yuan, J.; Love, D.J.; Erkip, E.; Niyato, D. A primer on rate-splitting multiple access: Tutorial, myths, and frequently asked questions. *arXiv* **2022**, arXiv:2209.00491.
26. Yang, J.; Ulukus, S. Delay-minimal transmission for average power constrained multi-access communications. *IEEE Trans. Wirel. Commun.* **2010**, *9*, 2754–2767. [CrossRef]
27. Yeh, E. Delay-optimal rate allocation in multiaccess communications: A cross-layer view. In Proceedings of the IEEE Workshop on Multimedia Signal Processing, St.Thomas, VI, USA, 9–11 December 2002; pp. 404–407.
28. Goyal, M.; Kumar, A.; Sharma, V. Optimal cross-layer scheduling of transmissions over a fading multiaccess channel. *IEEE Trans. Inf. Theory* **2008**, *54*, 3518–3537. [CrossRef]
29. Koseoglu, M. Lower bounds on the LTE-A average random access delay under massive M2M arrivals. *IEEE Trans. Commun.* **2016**, *64*, 2104–2115. [CrossRef]
30. Awuor, F.M.; Wang, C.-Y. Massive machine type communication in cellular system: A distributed queue approach. In Proceedings of the IEEE International Conference on Communications, Kuala Lumpur, Malaysia, 22–27 May 2016; pp. 1–7.
31. Zhao, X.; Chen, W. Non-orthogonal multiple access for delay-sensitive communications: A cross-layer approach. *IEEE Trans. Commun.* **2019**, *67*, 5053–5068. [CrossRef]
32. Ding, Z.; Fan, P.; Poor, H.V. Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions. *IEEE Trans. Veh. Technol.* **2015**, *65*, 6010–6023. [CrossRef]
33. Sedaghat, M.A.; Müller, R.R. On user pairing in uplink NOMA. *IEEE Trans. Wirel. Commun.* **2018**, *17*, 3474–3486. [CrossRef]
34. Salehi, M.; Tabassum, H.; Hossain, E. Accuracy of distance-based ranking of users in the analysis of NOMA systems. *IEEE Trans. Commun.* **2019**, *67*, 5069–5083. [CrossRef]
35. Zhao, X.; Chen, W. Delay optimal non-orthogonal multiple access with joint scheduling and superposition coding. In Proceedings of the IEEE Global Communications Conference, Singapore, 4–8 December 2017; pp. 1–6.
36. Liu, L.; Sheng, M.; Liu, J.; Dai, Y.; Li, J. Stable throughput region and average delay analysis of uplink NOMA systems with unsaturated traffic. *IEEE Trans. Commun.* **2019**, *67*, 8475–8488. [CrossRef]
37. Sheng, M.; Jiao, W.; Wang, X.; Liu, G. Effect of power control on performance of users in an interference-limited network with unsaturated traffic. *IEEE Trans. Veh. Technol.* **2016**, *66*, 2740–2755. [CrossRef]
38. Zhong, Y.; Quek, T.Q.; Ge, X. Heterogeneous cellular networks with spatio-temporal traffic: Delay analysis and scheduling. *IEEE J. Sel. Areas Commun.* **2017**, *35*, 1373–1386. [CrossRef]
39. Xu, C.; Wu, M.; Xu, Y.; Fang, Y. Uplink low-power scheduling for delay-bounded industrial wireless networks based on imperfect power-domain NOMA. *IEEE Syst. J.* **2020**, *14*, 2443–2454. [CrossRef]
40. Nasfi, R.; Chorti, A. Performance analysis of the uplink of a two user NOMA network under QoS delay constraints. In Proceedings of the IEEE International Conference on Ubiquitous and Future Networks, Zagreb, Croatia, 2–5 July 2019; pp. 526–528.
41. Xiao, C.; Zeng, J.; Liu, B.; Su, X.; Wang, J. Cross-layer power control for uplink NOMA in IoT applications with statistical delay constraints. In Proceedings of the IEEE Global Communications Conference, Abu Dhabi, United Arab Emirates, 9–13 December 2018; pp. 1–7.
42. Bello, M.; Yu, W.; Pischella, M.; Chorti, A.; Fijalkow, I.; Musavian, L. Flexible multiple access enabling low-latency communications: Introducing NOMA-R. *arXiv* **2020**, arXiv:2001.10637.
43. Li, A.; Chen, X.; Jiang, H. Contention based uplink transmission with NOMA for latency reduction. In Proceedings of the IEEE Vehicular Technology Conference, Sydney, NSW, Australia, 4–7 June 2017; pp. 1–6.

44. Seo, J.-B.; Jung, B.C.; Jin, H. Performance analysis of NOMA random access. *IEEE Commun. Lett.* **2018**, *22*, 2242–2245. [CrossRef]
45. Zhai, D.; Zhang, R.; Cai, L.; Yu, F.R. Delay minimization for massive internet of things with non-orthogonal multiple access. *IEEE J. Sel. Top. Signal Process.* **2019**, *13*, 553–566. [CrossRef]
46. Park, T.; Lee, G.; Saad, W. Message-aware uplink transmit power level partitioning for non-orthogonal multiple access (NOMA). In Proceedings of the IEEE Global Communications Conference, Abu Dhabi, United Arab Emirates, 9–13 December 2018; pp. 1–6.
47. Choi, M.; Kim, J.; Moon, J. Dynamic power allocation and user scheduling for power-efficient and delay-constrained multiple access networks. *IEEE Trans. Wirel. Commun.* **2019**, *18*, 4846–4858. [CrossRef]
48. Sreya, G.; Saigadha, S.; Goutam, M.P.D.D.; S, D.H. Adaptive rate NOMA for cellular IoT networks. *Proc. IEEE Wirel. Commun. Lett.* **2021**, *11*, 478–482. [CrossRef]
49. Jinho, G. Opportunistic NOMA for uplink short-message delivery with a delay constraint. *IEEE Trans. Wirel. Commun.* **2020**, *19*, 3727–3737.
50. Schiessl, S.; Sebastian, M.; Skoglund, M.; Gross, J. NOMA in the uplink: Delay analysis with imperfect CSI and finite-length coding. *IEEE Trans. Wirel. Commun.* **2020**, *19*, 3879–3893. [CrossRef]
51. Zeng, J.; Xiao, C.; Li, Z.; Ni, W.; Liu, R. Dynamic Power Allocation for Uplink NOMA With Statistical Delay QoS Guarantee. *IEEE Trans. Wirel. Commun.* **2021**, *20*, 8191–8203. [CrossRef]
52. Kleinrock, L. *Queuing Systems, Volume I: Theory*; Wiley: New York, NY, USA, 1975.
53. Kleinrock, L. *Queuing Systems, Volume II: Computer Applications;* Wiley: New York, NY, USA, 1975.
54. Chan, W.; Lu, T.-C.; Chen, R.-J. Pollaczek-Khinchin formula for the M/G/1 queue in discrete time with vacations. *IEEE Proc. Comput. Digit. Tech.* **1997**, *144*, 222–226. [CrossRef]