Dynamic Power Allocation and User Scheduling for Power-Efficient and Delay-Constrained Multiple Access Networks

Minseok Choi, *Member, IEEE*, Joongheon Kim, *Senior Member, IEEE*, and Jaekyun Moon, *Fellow, IEEE*

***Abstract*—In this paper, we propose a joint dynamic power control and user pairing algorithm for power-efficient and delay-constrained hybrid multiple access systems. In a hybrid multiple access system, user pairing determines whether the transmitter serves as a certain user by orthogonal multiple access (OMA) or non-orthogonal multiple access (NOMA). The proposed optimization framework minimizes the long-term time-average transmit power expenditure while reducing the queuing delay and guaranteeing the minimum time-average data rates. The proposed technique observes both channel and queue state information and adjusts queue backlogs to avoid an excessive queueing delay by appropriate user pairing and power allocation. Furthermore, the flexible use of resources is captured in the proposed algorithm by employing NOMA. The data-intensive simulation results show that the proposed scheme for power allocation and user scheduling achieves a balance among multiple performance goals, i.e., power efficiency, queueing delay, and data rate.**

***Index Terms*—Delay-constrained networks, power-efficient networks, non-orthogonal multiple access (NOMA), Internet of Things (IoT), power allocation, user scheduling.**

I. Introduction

THE fifth-generation (5G) wireless networks are expected to offer high spectral efficiency, improved reliability, massive connectivity, and low end-to-end (E2E) latency [1]. With the proliferation of smart devices, particularly in the Internet of Things (IoT) network, the system should not only provide sufficiently high system throughput and delay-constrained services, but also support machine type communications on a massive scale and/or device-to-device (D2D) networks [2]. Therefore, power-efficiency becomes critical when the battery-powered small IoT device has a role of transmitting information to other devices [3], [4]. Further, flexibility is also important for communications among heterogeneous machine type devices while meeting a variety of quality of service (QoS) requirements [5]. Many researchers have studied a myriad of technical issues related to the trends mentioned above.

The delay-constrained communications have been a major challenge and interest for a long time in various wireless networks. Given a delay constraint, the tradeoff between reliability and delay has been studied [6], and throughput analysis also has been performed [7]. With the respect of this tradeoff, the packet delay can be reduced by adjusting the transmission policy [8]. The E2E delay consists of uplink (UL)/downlink (DL) transmission delays and the queueing delay [9], [10], and a short frame structure reduces UL/DL transmission durations [12]. Meanwhile, the analysis of deterministic queueing delay is very difficult because of the fact that queue dynamics in medium access control (MAC) are influenced by the randomness of time-varying channels and the stochastic geometry in the physical (PHY) layer.

As described in [13], the *effective capacity* link-layer model can be used to define the statistical delay requirement. Based on the *effective capacity* model, cross-layer transmission designs for achieving queueing delay requirements have been investigated in [14], [15]. Further, based on Little's theorem [16], which establishes that the time-average queueing delay is proportional to the average queue backlog, delay-constrained scheduling has been proposed in [17] by pursuing stability of queuing systems of delay-constrained users. In this respect, dynamic resource allocation and scheduling policies which reduce the queueing delay by limiting time-average queue backlogs have been actively researched in [8], [18]–[20].

Since there exists a fundamental power-delay tradeoff as studied in [21], [22], power-efficiency is also critical for delay-constrained communications, especially where a massive number of devices are battery powered [23]. Energy-efficient resource allocations and scheduling policies for delay-constrained communications have been studied in [24]–[26], and a delay-optimal scheduling policy for power-constrained transmission has been proposed in [27]. In addition, system throughput maximization subject to the queueing delay constraint was addressed in [28]. Furthermore, the tradeoff between energy and delay depending on changes in the network state distribution was discussed in [29], [30] based on a stochastic network optimization framework.

Manuscript received December 23, 2018; revised May 8, 2019; accepted July 15, 2019. Date of publication July 26, 2019; date of current version October 9, 2019. This work was supported by Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) under Grant 2017-0-00068–A Development of Driving Decision Engine for Autonomous Driving using Driving Experience Information. The associate editor coordinating the review of this paper and approving it for publication was C. Zhong. *(Corresponding authors: Joongheon Kim; Jaekyun Moon.)*

M. Choi is with the Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA 90089 USA (e-mail: choimins@usc.edu).

J. Kim is with the School of Electrical Engineering, Korea University, Seoul 02841, South Korea (e-mail: joongheon@korea.ac.kr).

J. Moon is with the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, South Korea (e-mail: jmoon@kaist.edu).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2019.2929809

Moreover, as a massive number of various devices is deployed in the network, orthogonal multiple access (OMA) is no longer able to maximize resource efficiency and to serve all devices simultaneously. In order to overcome this issue, non-orthogonal multiple access (NOMA) has been actively researched as one of the promising methods for the efficient and flexible use of both energy and the spectrum, as well as for system throughput improvements [31]. Power-multiplexing NOMA provides a better system throughput than OMA with the ideal successive interference cancellation (SIC) [32]. In addition, NOMA has the advantage of allowing massive connectivity for IoT services [33], and NOMA in short packet transmissions for achieving low latency has been discussed in [34], [35]. Cooperative NOMA schemes assisted by D2D communications have been also studied in [36], [37].

Since all users would not be served by NOMA as a result of the high complexity of SIC, hybrid multiple access (MA), which allows for the coexistence of NOMA and OMA, has been considered for next-generation communication systems. Representatively, multi-user superposition transmission (MUST) has been adopted by the 3rd Generation Partnership Project (3GPP) for 5G networks, which employs both power-domain NOMA and orthogonal frequency division multiple access (OFDMA) [38]. For the use of hybrid MA, user scheduling and resource allocation are very critical issues. In [39]–[41], user pairing schemes for NOMA signaling have been studied, and joint resource allocations in NOMA systems have been considered in [42]–[45]. However, power-efficiency and low latency were not considered in [39]–[44]. The authors of [45] and [46] proposed power-efficient resource allocation policies, but they did not consider user pairing and delay problems.

This paper proposes the dynamic policy for user pairing and power control to maximize power efficiency while achieving delay constraints as well as sufficient reliability in hybrid MA. In particular, the long-term average data rate is considered as a user QoS requirement for sufficient reliability. In addition, user scheduling and flexible use of resources are also captured in the proposed technique. The main contributions of the proposed technique can be summarized as follows:

- This paper constructs the stochastic network optimization framework for the transmission scheme of power-efficient and delay-constrained communications, which adaptively operates depending on time-varying channel and queue states. The proposed framework focuses on reducing the queueing delay, which is a main factor in the E2E delay.
- This paper contributes to delay-constrained systems based on NOMA. The proposed transmission scheme utilizes the advantage of NOMA over OMA to increase the data rate for reducing the queueing delay. Further, users enable to utilize resources flexibly by employing NOMA.
- Different from the existing power-efficient resource allocations [45], [46], the proposed resource allocation and user scheduling not only maximize power-efficiency, but also guarantee limited queueing delays and sufficiently large time-average data rates for all users.
- Data-intensive simulation results show that the proposed technology can achieve a balance among multiple performance requirements, i.e., power expenditure, queueing delay and data rate, on the basis of the short packet structure [12].

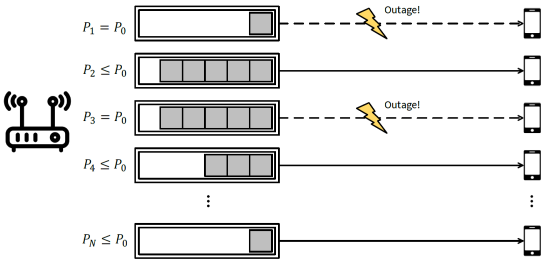

Fig. 1. System architecture for OMA.

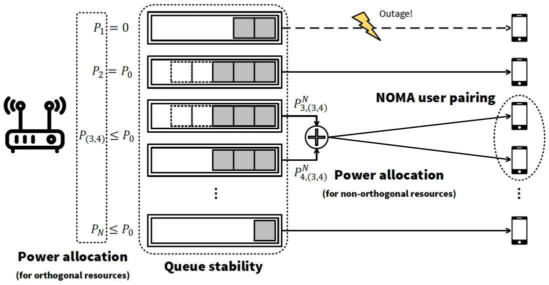

Fig. 2. System architecture for hybrid MA.

The rest of the paper is organized as follows. The hybrid MA system and queue model are described in Section II. In Section III, we formulate the joint optimization problem of user pairing and power allocation in hybrid MA. The optimal power allocation rule with fixed user pairing is proposed in Section IV-B, and the matching algorithm for user pairing is presented in Section V. Simulation results are shown in Section VI, and Section VII concludes the paper.

II. System Model

A. Hybrid Multiple Access Model

This paper considers hybrid MA for power-efficient and delay-constrained IoT networks. Let a transmitter serve N users by employing either OMA only or NOMA, as shown in Figs. 1 and 2, respectively. The transmitter is deployed with N queues in which data packets are waiting for transmissions to N users, respectively. The data packets for user n are accumulated in queue n. The data transmission is performed in each discrete time slot, i.e., $t = 1, 2, \cdots$. Suppose that each transmitter queue has a power budget of P_0, and the transmit power for user n is P_n; therefore, $0 \leq P_n \leq P_0$. Here, we assume the maximum transmit power constraint for each link or subchannel rather than the sum power constraint. Since the individual power constraint per link is stricter than the sum power constraint, the power allocation satisfying per-link power constraints can also always satisfy the sum power constraint by using an appropriate value of P_0, even though it yields a loss of power efficiency. Accordingly, if individual power constraints are considered, then there is no need of the sum-power constraint. In addition, the individual power constraints are also beneficial to reduce the peak-to-average power ratio which is a critical issue especially in multicarrier systems.

The Rayleigh fading channel is assumed for communication links from the transmitter to users. Denote the channel of user n with h_n. The path loss model is $35.3 + 37.6\ln(d_n)$, where d_n is the distance between the transmitter and user n, and the fast fading component has a complex Gaussian distribution, i.e., $CN(0, 1)$. The channel is assumed to be static during each discrete time slot. Let R_n be the data rate for user n, and denote ρ_n as a threshold of user n's instantaneous data rate.

When $R_n < \rho_n$, an outage event occurs at user n. In addition, η_n represents the long-term average data rate as a QoS requirement for user n, and the QoS constraint can be written as

$$\eta_n \leq \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} R_n(t). \tag{1}$$

B. Transmitter Queue Model

In general, the transmitter queue model has its own arrival and departure processes. When departures are less frequent than arrivals, the queue backlog grows. For each user $n \in \{1, \cdots, N\}$, the queue dynamics in each discrete time slot $t \in \{0, 1, \cdots, \}$ can be represented as follows:

$$Q_n(t+1) = \max\{Q_n(t) - \mu_n(t), 0\} + \lambda_n(t) \tag{2}$$
$$Q_n(0) = 0, \tag{3}$$

where $Q_n(t)$, $\lambda_n(t)$, and $\mu_n(t)$ stand for the queue backlog and the arrival and departure processes of user n at time t, respectively. The queue states are updated in each time slot t. In this paper, the interval of each slot is assumed to be the channel coherence time, τ_c.

In this paper, queue backlog $Q_n(t)$ represents data bits waiting to be transmitted to user n.

$\lambda_n(t)$ and $\mu_n(t)$ semantically mean arrived and transmitted bits at slot t, respectively. Simply, suppose that $\lambda_n(t)$ is randomly generated for all $n \in \{1, \cdots, N\}$. On the other hand, $\mu_n(t)$ obviously depends on the data rate of user n as follows:

$$\lambda_n(t) = a_n(t) \cdot u \tag{4}$$
$$\mu_n(t) = \mathcal{I}\{R_n(P_n, \Psi_n, t) \geq \rho_n\} \cdot R_n(P_n, \Psi_n, t) \cdot \tau_c, \tag{5}$$

where $a_n(t)$ is an i.i.d. uniform random variable, i.e., $a_n(t) \sim \mathcal{U}\{\lambda_{\min}, \lambda_{\max}\}$, indicating the number of data packets that have arrived in queue n at time t. Also, u is the packet size in bits, and Ψ_n represents the index of the user paired with user n. If OMA is employed for user n, then $\Psi_n = n$, whereas $\Psi_n = m$ for $m \neq n$ means that users n and m are paired for NOMA. $R_n(P_n, \Psi_n, t)$ is the data rate of user n when transmit power is P_n and user n is paired with user Ψ_n at time t. $\mathcal{I}(.)$ is the indicator function, and $\mathcal{I}\{R_n(P_n, \Psi_n, t) \geq \rho_n\}$ is 0 if the outage event occurs at user n; otherwise, it is 1.

Remark: If τ_c is too long, it is better to update the power allocation and user pairing more frequently than channel variations. Consider a transmitter queue that is almost empty so that there is no worry about excessive queueing delays. In this case, the transmitter usually consumes a small amount of power to improve power–efficiency. However, if this situation persists for a long time, as indicated by τ_c, packets will accumulate in the queue and queueing delays will increase. Therefore, several updates in regard to power allocation and user pairing are required over the time interval of τ_c, i.e., the interval of each time slot should be smaller than τ_c.

C. Delay Constraint and Queue Stability

Denote the E2E delay bound with $D_{\max}$. $D_{\max}$ mainly consists of UL/DL transmission delays and the queueing delay [11]. For delay-constrained communications, a small packet structure is preferred because UL/DL transmission durations can be reduced. Then, the summation of UL/DL durations becomes identical to the transmit time interval (TTI), denoted by T_t [12]. Therefore, the margin of the queueing delay is $D^q_{\max} = D_{\max} - T_t$, i.e., data transmission is successful only when the queueing delay is smaller than $D^q_{\max}$. However, making the instantaneous queueing delay bounded to a deterministic value is very difficult due to time-varying channel environments and dynamic transmissions.

To this end, this paper focuses on limiting the time-average queueing delay. According to Little's theorem [16], the time-average queueing delay is proportional to the average queue length. In addition, the Lyapunov optimization theory [47] proved that the time-average queue backlogs can be limited by pursuing strong stability of a queue defined as

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T} \mathbb{E}[Q(t)] < \infty. \tag{6}$$

Based on the Lyapunov optimization theory, the upper bound on the time-average queue length is also derived by using the algorithm which minimizes the Lyapunov drift [47] and finally the delay constraint can be satisfied by achieving queue stability in (6). In this respect, many delay-constrained transmission policies which limit the queueing delay by pursuing the queue stability have been proposed in [8], [18]–[20]. In this paper, simulation results in Section VI show that the queueing delay can be reduced by ensuring (6), i.e., strong stability of the queueing system.

D. Problem Scenario

In this subsection, the example scenario of the proposed optimization problem is presented. Fig. 1 shows the system architecture employing OMA only. For example, many packets are accumulated in queue 2 and queue 3; therefore, these links have a risk of excessive queueing delays. The transmitter can reduce queue backlogs by increasing transmission rates of these links. Simply, link 2 can consume more power to increase its transmission rate and to reduce queue backlogs. On the other hand, link 3 experiences the outage event even with the maximum transmit power in Fig. 1; therefore, it cannot avoid the excessive queueing delay. In this case, NOMA can help link 3 to satisfy the delay constraint. By sharing the frequency resource with another link, link 3 can utilize larger bandwidth than before, but should handle interference from another link. Since NOMA is well-known to improve system throughput compared to OMA with the ideal SIC, NOMA has a potential to increase the data rate of link 3 and finally to satisfy its delay constraint.

The ideal SIC cannot be employed in practice, however, the appropriate user pairing and power allocations are necessary for NOMA. In Fig. 2, user 3 and user 4 are paired for NOMA signaling and link 3 is no longer in outage. In this scenario, the data rate of link 4 could be degraded, but it is fine because queue 4 has enough queue backlogs to endure the excessive queueing delay for a while. Similarly, the link outage occurs at user 1 whose is almost empty so that there is no need to worry about the long queueing delay.

In this case, the transmitter can deactivate link 1 to save power consumption, as shown in Fig. 2. In this way, the system dynamically adjusts transmission rates of all links by controlling transmit power and employing NOMA to achieve both low queueing delay and high power efficiency.

Since this model determines link deactivation by allocating no transmit power, as shown by link 1 in Fig. 2, user scheduling is also performed in the system model. If some links are determined not to active, then their resources would be shared by other links in terms of resource efficiency. The flexible resource use among users can be also realized in the presented system model by pairing two links for NOMA signaling and allocating no power to one of them. It means that the orthogonal resource of the user with no transmit power is occupied by another paired user.

The main issues associated with hybrid MA are summarized in Fig. 2. First, the queueing delay should be reduced to satisfy the delay constraint by limiting the time-average queue backlogs, i.e., achieving stability of the queueing system. Second, when NOMA is utilized, user pairing problem, i.e., which user is better to be paired for NOMA with certain user who needs to raise the data rate, arises. Finally, power allocations for both OMA and NOMA users should be jointly determined with the user pairing problem. In this paper, only a two-user NOMA scenario is considered because small devices in the IoT network are difficult to handle the high computational complexity of SIC processes for the multi-user NOMA scenario. Further, as the number of power-multiplexed users increases, the required power budget becomes larger to provide reliable signal-to-interference-plus-noise ratios (SINRs) to all NOMA users. However, a small battery-powered device generally has a limited power budget.

III. Joint Optimization Problem Formulation for User pairing and Power Allocation in Hybrid Multiple Access

This paper pursues both high power-efficiency and low queueing delay. In addition, the long-term average data rate is considered as user's the QoS requirement. The data rate of each user depends on its transmit power and MA scheme, i.e., OMA or NOMA. If NOMA is employed for certain user n, then its data rate is determined by which user is paired with user n. Therefore, we denote the data rate of user n at time t by $R_n(P_n(t), \Psi_n(t), t)$ to describe the dependencies on power allocation and user pairing.

The joint optimization problem to find the optimal power allocation and user pairing can be formulated as follows:

$$\begin{aligned}
&\{\mathbf{P}^*(t), \mathbf{\Psi}^*(t)\} \\
&\quad = \underset{\mathbf{P},\mathbf{\Psi}}{\arg\min} \sum_{n\in\mathcal{N}} \mathbb{E}[P_n(t)] && (7)\\
&\quad\quad \text{s.t.} \lim_{t\to\infty} \frac{1}{t} \sum_{t'=0}^{t} \mathbb{E}[Q_n(t')] < \infty, \quad \forall n \in \mathcal{N} && (8)\\
&\quad\quad\quad \lim_{t\to\infty} \frac{1}{t} \sum_{t'=0}^{t-1} \mathbb{E}[\tilde{R}_l(P_l(t'), \Psi_l(t'), t')] \geq \eta_l, \\
&\quad\quad\quad \forall l \in \mathcal{N}_s \subseteq \mathcal{N} && (9)\\
&\quad\quad\quad 0 \leq P_n(t) \leq P_0, \quad \forall n \in \mathcal{N} && (10)\\
&\quad\quad\quad \Psi_n(t) \in \mathcal{N}, \quad \forall n \in \mathcal{N} && (11)\\
&\quad\quad\quad \Psi_n(t) = m \text{ if and only if } \Psi_m(t) = n, && (12)
\end{aligned}$$

where $\tilde{R}_n(.) = \mathcal{I}\{R_n(.) \geq \rho_n\} \cdot R_n(.)$, $\mathcal{N} = \{1, \cdots, N\}$ is the user index set, and $\mathcal{N}_s$ is the subset of $\mathcal{N}$. $\mathbf{P}^*(t)$ and $\mathbf{\Psi}^*(t)$ denote vectors of the optimal power allocations and user pairings, i.e., $P_n^*(t)$ and $\Psi_n^*(t)$, for all $n \in \mathcal{N}$, respectively. The constraint (8) represents strong stability of the queueing system, which makes queue backlogs upper bounded. In addition, the minimum time-average data rate η_l of user l for $l \in \mathcal{N}_s$ is guaranteed as one of the QoS requirements stipulated by constraint (9). Again, we assume the individual power constraint for each link in (10) which is stricter than the sum power constraint and beneficial to reducing the peak-to-average power ratio in the multicarrier system.

The problem in (7)–(10) can be solved by the theory of Lyapunov optimization [47]. We first transform the inequality constraint (9) into the form of queue stability. Specifically, first define the virtual queues $Z_l(t)$ for all $l \in \mathcal{N}_s$ with the update equation:

$$Z_l(t+1) = \max\{Z_l(t) + \eta_l - \tilde{R}_l(P_l(t), \Psi_l(t), t), 0\}. \quad (13)$$

The strong stability of the virtual queue $Z_l(t)$ pushes the average of $\tilde{R}_l(P_l(t), \Psi_l(t), t)$ to be close to the QoS guarantee η_l.

Let $\mathbf{Q}(t)$ and $\mathbf{Z}(t)$ denote the column vectors of $Q_n(t)$ and $Z_l(t)$ for $n \in \mathcal{N}$ and $l \in \mathcal{N}_s$ at time t, respectively, and let $\mathbf{\Theta}(t) = [\mathbf{Q}(t)^T, \mathbf{Z}(t)^T]^T$ be a concatenated vector of actual and virtual queue backlogs. Define the quadratic Lyapunov function $L[\mathbf{\Theta}(t)]$ as follows:

$$L[\mathbf{\Theta}(t)] = \frac{1}{2} \sum_{n\in\mathcal{N}} Q_n(t)^2 + \frac{1}{2} \sum_{l\in\mathcal{N}_s} Z_l(t)^2. \quad (14)$$

Then, let $\Delta(t)$ be a conditional quadratic Lyapunov function that can be formulated as $\mathbb{E}[L[\mathbf{\Theta}(t+1)] - L[\mathbf{\Theta}(t)] | \mathbf{\Theta}(t)]$, i.e., the drift on t. The dynamic policy is designed to solve the optimization problem in (7)–(10) by observing the current queue state, $\mathbf{\Theta}(t)$, and determining power allocation $\mathbf{P}(t)$ and user pairing $\mathbf{\Psi}(t)$ in such a way as to minimize the upper bound on the *drift-plus-penalty* [47]:

$$\Delta(\mathbf{\Theta}(t)) + V\mathbb{E}\Big[\sum_{n\in\mathcal{N}} P_n(t)\Big|\mathbf{\Theta}(t)\Big], \tag{15}$$

where V is a positive constant of the policy that affects the tradeoff between the power efficiency and queueing delay.

First, find the upper bound on the change in the Lyapunov function.

$$\begin{aligned}
&L[\mathbf{\Theta}(t+1)] - L[\mathbf{\Theta}(t)]\\
&\quad= \frac{1}{2}\sum_{n\in\mathcal{N}}\Big[Q_n(t+1)^2 - Q_n(t)^2\Big]\\
&\qquad+\frac{1}{2}\sum_{l\in\mathcal{N}_s}\Big[Z_l(t+1)^2 - Z_l(t)^2\Big] \qquad (16)\\
&\quad\le \frac{1}{2}\sum_{n\in\mathcal{N}}[\lambda_n(t)^2 + \mu_n(t)^2] + \sum_{n\in\mathcal{N}} Q_n(t)(\lambda_n(t) - \mu_n(t))\\
&\qquad+\frac{1}{2}\sum_{l\in\mathcal{N}_s}(\eta_l - \tilde{R}_l(P_l(t), \Psi_l(t), t))^2\\
&\qquad+\sum_{l\in\mathcal{N}_s} Z_l(t)(\eta_l - \tilde{R}_l(P_l(t), \Psi_l(t), t)). \qquad (17)
\end{aligned}$$

Then, the upper bound on the conditional Lyapunov drift is given by

$$\begin{aligned}
\Delta(\Theta(t)) \le C &+ \sum_{n\in\mathcal{N}}\mathbb{E}\Big[Q_n(t)(\lambda_n(t) - \mu_n(t))\Big]\\
&+\sum_{l\in\mathcal{N}_s}\mathbb{E}\Big[Z_l(t)(\eta_l - \tilde{R}_l(P_l(t), \Psi_l(t), t))\Big]. \qquad (18)
\end{aligned}$$

where we assume that departure and arrival rates are bounded, and C is a constant such that $\frac{1}{2}\sum_{n\in\mathcal{N}}\mathbb{E}[\lambda_n(t)^2 + \mu_n(t)^2] + \frac{1}{2}\sum_{l\in\mathcal{N}_s}\mathbb{E}[(\eta_l - \tilde{R}_l(P_l(t), \Psi_l(t), t))^2] \le C$. According to (15), minimizing the upper bound on the drift-plus-penalty is consistent with minimizing

$$\begin{aligned}
\mathbb{E}\Bigg[V\sum_{n\in\mathcal{N}} P_n(t) &- \sum_{n\in\mathcal{N}} Q_n(t)\mu_n(t)\\
&- \sum_{l\in\mathcal{N}_s} Z_l(t)\tilde{R}_l(P_l(t), \Psi_l(t), t)\Big|\mathbf{\Theta}(t)\Bigg], \qquad (19)
\end{aligned}$$

because $\lambda_n(t)$ is not controllable and all values of η_l for $l \in \mathcal{N}_s$ are constants.

We now use the concept of opportunistically minimizing the expectations and specifically go after the following drift-plus-penalty problem:

$$\begin{aligned}
\{\mathbf{P}^*(t), \mathbf{\Psi}^*(t)\} &= \underset{\mathbf{P},\mathbf{\Psi}}{\arg\min}\ \mathcal{M}(\mathbf{P}(t), \mathbf{\Psi}(t)) \qquad (20)\\
&\text{s.t. } (10), (11), (12) \qquad (21)
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{M}(\mathbf{P}(t), \mathbf{\Psi}(t)) &= \sum_{n\in\mathcal{N}} \mathcal{M}_n(P_n(t), \Psi_n(t)) \qquad (22)\\
&= V\sum_{n\in\mathcal{N}} P_n(t) - \sum_{n\in\mathcal{N}} Q_n(t)\mu_n(t)\\
&\quad- \sum_{l\in\mathcal{N}_s} Z_l(t)\tilde{R}_l(P_l, \Psi_l, t). \qquad (23)
\end{aligned}$$

Since there are very many possible combinations of user pairing, it is difficult to exhaustively minimize the optimization metric of (23). Therefore, we first find the optimal power allocation depending on the fixed user pairing policy. Then, several pairs of two users are generated for NOMA to minimize the optimization metric of (23) based on the matching theory.

IV. Optimal Power Allocation for Hybrid Multiple Access

For simplicity, notations for the dependency of all parameters on t are omitted in this section, because the optimal power allocation depends only on channel state information (CSI) and queue state information (QSI) at current time t. Therefore, $R_n(P_n(t), \Psi_n(t), t) = R_n(P_n, \Psi_n)$ in this section.

A. Optimal Power Allocation for Orthogonal Multiple Access

First, the power allocation rule for OMA users is presented. The data rate of user n, which employs OMA, is given by

$$R_n(P_n, n) = \frac{\Phi\mathcal{B}}{N}\log_2\Big(1 + N\Gamma_n P_n\Big), \tag{24}$$

where $\Gamma_n = \frac{|h_n|^2}{\mathcal{B}N_0}$, N_0 is the single-sided noise spectral density, and $\mathcal{B}$ is the bandwidth. $\Phi \in (0, 1]$ represents the degradation coefficient of the channel capacity due to the finite blocklength codes appropriate for the short packet structure [11], [48]. Note that the bandwidth is equally allocated to N users, and $\Psi_n = n$ for all $n \in \mathcal{N}$ in OMA. The power interval for avoiding the outage, i.e., $R_n(P_n, n) \ge \rho_n$, can be obtained by

$$P_n \ge P_{\text{th}}^O = \frac{2^{N\rho_n/\Phi\mathcal{B}} - 1}{N\Gamma_n}. \tag{25}$$

If $P_n \le P_{\text{th}}^O$, then $\tilde{R}_n(P_n, n) = 0$.

Remark: Since Shannon capacity assumes channel codes of infinite length, it is not appropriate to directly apply Shannon capacity to low-latency communications with the short packet structure. The authors of [48], [49] obtained the channel capacity with finite blocklength codes in a variety of channel models. The definition of the data rate in (24) is the normal approximation of the data rate with the finite blocklength code [48]. When the decoding error probability is 10^{-6} and SNR is larger than 20 dB in AWGN channel, $\Phi = 0.9$ can be used. Therefore, we assume that the channel code is used whose decoding error probability is lower than 10^{-6}. Since the channel is static during each discrete time slot, this fading channel can be viewed as an AWGN channel with channel gain $|h_n|^2$ at user n, and all simulation results are obtained with received SNR larger than 20 dB in Section VI.

Since each OMA link is orthogonal to other links, each user's data rate $R_n(P_n, n)$ is independent of each other; therefore, the optimization problem in (20)–(21) can be solved by independently minimizing $\mathcal{M}_n(P_n, \Psi_n)$. When $\Psi_n = n$, let $\mathcal{M}_n^O(P_n) = \mathcal{M}_n(P_n, \Psi_n)$. Therefore, the optimization problem in (20)–(21) can be transformed into

$$\begin{aligned}
P_n^O &= \underset{P_n}{\arg\min}\ \mathcal{M}_n^O(P_n) \qquad (26)\\
&\text{s.t. } 0 \le P_n \le P_0, \qquad (27)
\end{aligned}$$

where

$$\mathcal{M}_n^O(P_n) = V \cdot P_n - Q_n \cdot \tilde{R}_n(P_n, n) \cdot \tau - Z_n \cdot \tilde{R}_n(P_n, n) \cdot \mathcal{I}\{n \in \mathcal{N}_s\}. \quad (28)$$

Theorem 1 provides the solution of the optimization problem in (26)–(27).

Theorem 1: The optimal power allocation of the problem in (26)–(27) is given by

$$P_n^O = \begin{cases} P_0, & \text{if } P_{\text{th}}^O \le P_0 \le P_n^* \ \& \ \mathcal{M}_n(P_0) < 0 \\ P_n^*, & \text{else if } P_{\text{th}}^O \le P_n^* < P_0 \\ P_{\text{th}}^O, & \text{else if } P_n^* < P_{\text{th}}^O \le P_0 \ \& \ \mathcal{M}_n(P_{\text{th}}^O) < 0 \\ 0, & \text{otherwise,} \end{cases} \quad (29)$$

where $P_n^* = \frac{\Phi\mathcal{B}(\tau Q_n + Z_n \cdot \mathcal{I}\{n \in \mathcal{N}_s\})}{NV \ln 2} - \frac{1}{N\Gamma_n}$.

Proof: Assume that $P_n \ge P_{\text{th}}^O$, i.e., the outage event does not occur. Then, differentiating (28) by P_n,

$$\frac{\mathrm{d}\mathcal{M}_n^O}{\mathrm{d}P_n} = V - \frac{\Phi\mathcal{B}(\tau Q_n + \tilde{Z}_n)}{\ln 2} \cdot \frac{\Gamma_n}{N\Gamma_n P_n + 1}, \quad (30)$$

where $\tilde{Z}_n = Z_n \cdot \mathcal{I}\{n \in \mathcal{N}_s\}$, and the local minimizer P_n^* is obtained from $\frac{\mathrm{d}\mathcal{M}_n^O}{\mathrm{d}P_n} = 0$, i.e.,

$$P_n^* = \frac{\Phi\mathcal{B}(\tau Q_n + \tilde{Z}_n)}{NV \ln 2} - \frac{1}{N\Gamma_n}. \quad (31)$$

Further, P_n^* is shown to be the global minimizer in the region of $P_n \ge P_{\text{th}}^O$ by

$$\frac{\mathrm{d}^2\mathcal{M}_n^O}{\mathrm{d}P_n^2} = \frac{N\Gamma_n^2 \Phi\mathcal{B}(\tau Q_n + \tilde{Z}_n)}{(N\Gamma_n P_n + 1)^2 \ln 2} > 0. \quad (32)$$

However, when $P_n < P_{\text{th}}^O$, $\mathcal{M}_n^O(P_n) = V \cdot P_n$; therefore, $P_n = 0$ is the minimizer and $\mathcal{M}_n^O(P_n) = 0$. If $P_{\text{th}}^O > P_0$, $P_n^O = 0$ always. Otherwise, i.e. when $P_{\text{th}}^O \le P_0$, the relative value of P_n^* to P_{th}^O and P_0 determines P_n^O.

When $P_{\text{th}}^O \le P_n^* < P_0$, P_n^* is still the global minimizer. However, when $P_0 \le P_n^*$, the minimizer in the interval of $[P_{th}^O, P_0]$ becomes P_0, but $P_n = 0$ is the minimizer in $[0, P_{th}^O]$. Therefore, if $\mathcal{M}_n^O(P_0) < 0$, P_0 is the global minimizer in $[0, P_0]$. Otherwise, $P_n^O = 0$, i.e., no power is allocated to user n.

In the case of $P_n^* < P_{\text{th}}^O$, the global minimizer is in the outage region. Then, P_{th}^O is the minimizer in the interval of $[P_{\text{th}}^O, P_0]$. Thus, if $\mathcal{M}_n^O(P_{\text{th}}^O) < 0$, $P_n^O = P_{\text{th}}^O$ becomes the minimizer in $[0, P_0]$, and if not, $P_n^O = 0$ is the solution. Finally, (29) is obtained. □

Remark: When the outage is expected at user n by observing CSI and QSI, the transmitter can save the power, i.e., $P_n = 0$. Further, when queue backlogs of Q_n and Z_n are small and that the second and third terms of (28) are small compared to the system parameter V, the link of user n is not scheduled to save the power, even though the link is not in the outage. Thus, it can be said that link scheduling is also performed by utilizing power allocations.

B. Optimal Power Allocation for Two-User Non-orthogonal Multiple Access

In this section, the optimal power allocation in the NOMA system is obtained for a given pair of user i and user j, i.e., $\Psi_i = j$ and $\Psi_j = i$. Assume that $|h_j|^2 > |h_i|^2$. For employing NOMA, the larger power is generally allocated to the user with a weaker channel condition. Throughout the paper, the user with the weaker channel who is not subjected to SIC and the user with the stronger channel who is necessary to perform SIC will be referred to the non-SIC user and SIC user, respectively. Let user i and user j be the non-SIC user and the SIC user, respectively, with the assumption of $P_i \ge P_j$. The data rates of the NOMA users are given by

$$R_i(P_i, \Psi_i = j) = \frac{2\Phi\mathcal{B}}{N} \log_2 \left(1 + \frac{N\Gamma_i P_i/2}{N\Gamma_i P_j/2 + 1}\right) \quad (33)$$

$$R_j(P_j, \Psi_j = i) = \frac{2\Phi\mathcal{B}}{N} \log_2(1 + N\Gamma_j P_j/2). \quad (34)$$

Suppose that signals for other users are orthogonally multiplexed with the NOMA signaling of user i and user j. Then, the power allocation problem for user i and user j can be formulated from the power allocation problem of (20)–(21) as follows:

$$\{P_{i,(i,j)}^N, P_{j,(i,j)}^N\} = \underset{P_i, P_j}{\arg\min}\ \mathcal{M}_{(i,j)}^N(P_i, P_j) \quad (35)$$

$$\text{s.t. } 0 \le P_i, P_j \le P_0, \quad (36)$$

where $\mathcal{M}_{(i,j)}^N(P_i, P_j) = \mathcal{M}_{i,(i,j)}^N(P_i) + \mathcal{M}_{j,(i,j)}^N(P_j)$, and $\mathcal{M}_{i,(i,j)}^N(P_i) = V \cdot P_i - (\tau Q_i + \tilde{Z}_i) \cdot \tilde{R}_i(P_i, \Psi_i = j)$, which is the optimization metric of user i when user i is paired with user j and $j \ne i$. $P_{j,(i,j)}^N$ represents the optimal transmit power for user j when user j is paired with user i for NOMA. However, $\mathcal{M}_{(i,j)}^N(P_i, P_j)$ is not concave; therefore, the optimization problem of (35)–(36) is not a convex problem. Therefore, the auxiliary variable $q = P_i + P_j$ is introduced, and $q \le 2P_0$ should be satisfied. Then, the problem of (35)–(36) can be resolved by solving two sequential subproblems.

The first subproblem is to find the power allocation for NOMA users with the fixed value of q, as formulated by

$$P_{j,(i,j)}^N = \underset{P_j}{\arg\min}\ g(P_j) \quad (37)$$

$$\text{s.t. } 0 \le P_j \le P_i \le P_0 \quad (38)$$

$$P_i + P_j = q, \quad (39)$$

where

$$g(P_j) = V \cdot q - (\tau Q_i + \tilde{Z}_i)\tilde{R}_i(P_j) - (\tau Q_j + \tilde{Z}_j)\tilde{R}_j(P_j). \quad (40)$$

Here, $\tilde{R}_i(.) = \mathcal{I}\{R_i(.) \ge \rho\}$, $R_i(P_j) = \frac{2\Phi\mathcal{B}}{N} \log_2 \left(\frac{N\Gamma_i q/2+1}{N\Gamma_i P_j/2+1}\right)$ and $R_j(P_j) = R_j(P_j, i)$ in (34).

The power intervals for avoiding the outage event at both NOMA users are considered to represent the solution to the subproblem of (37)–(39). $R_j(P_j, i) \ge \rho_j$ should be guaranteed for user j to avoid the outage, in other words, the transmit power should be

$$P_j \ge P_{j,(i,j)}^o = \frac{2}{N\Gamma_j}(2^{N\rho_j/2\Phi\mathcal{B}} - 1). \quad (41)$$

Similarly, user i can prevent the outage event when $R_i^N(P_i, j) \geq \rho_i$, and it corresponds to

$$P_j \leq P^o_{i,(i,j)} = \left[q - \frac{2}{N\Gamma_i} 2^{N\rho_i/2\Phi\mathcal{B}}\right] \cdot 2^{-N\rho_i/2\Phi\mathcal{B}}. \quad (42)$$

Let $\mathcal{O}_j = [0, P^o_{j,(i,j)}]$ and $\mathcal{O}_i = [P^o_{i,(i,j)}, q]$ denote the outage regions of user j and user i, respectively. When $P_j \in \mathcal{O}_j$ but $P_j \in \mathcal{O}_i^c \cap [0, P_0]$, the objective function of (40) is given by

$$g_{\mathcal{O}_j}(P_j) = V \cdot q - \frac{2\Phi\mathcal{B}(\tau Q_i + \tilde{Z}_i)}{N} \log_2\Big(\frac{N\Gamma_i q/2 + 1}{N\Gamma_i P_j/2 + 1}\Big). \quad (43)$$

Meanwhile, when $P_j \in \mathcal{O}_i$ but $P_j \in \mathcal{O}_j^2 \cap [0, P_0]$, the objective function of (40) becomes

$$g_{\mathcal{O}_i}(P_j) = V \cdot q - \frac{2\Phi\mathcal{B}(\tau Q_j + \tilde{Z}_j)}{N} \log_2\Big(1 + N\Gamma_j P_j/2\Big). \quad (44)$$

If both users i and j are in outage, i.e., $P_j \in \mathcal{O}_j \cup \mathcal{O}_i$, $g(P_j) = 0$.

Since $N\rho_i/2\Phi\mathcal{B} > 0$ always, $0 \leq P^o_{j,(i,j)}$ and $P^o_{i,(i,j)} \leq q$ are guaranteed. Then, Theorem 2 gives the solution to the subproblem of (37)–(39).

Theorem 2: Suppose that $1 < \frac{\tau Q_i + \tilde{Z}_i}{\tau Q_j + \tilde{Z}_j} < \frac{\Gamma_j}{\Gamma_i}$. When $P^o_{j,(i,j)} \leq P^o_{i,(i,j)}$, the optimal power allocation of the problem (37)-(39) is given by (46)-(51) if $g(P^N_{j,(i,j)}) < 0$, where $\bar{q} = \max(0, q - P_0)$ and

$$P_j^* = \frac{2}{N\Gamma_i\Gamma_j} \cdot \frac{\Gamma_j(\tau Q_j + \tilde{Z}_j) - \Gamma_i(\tau Q_i + \tilde{Z}_i)}{\tau Q_j + \tilde{Z}_j - \tau Q_i - \tilde{Z}_i}. \quad (45)$$

When $P^o_{j,(i,j)} > P^o_{i,(i,j)}$, the optimal power allocation is given by (52)–(54) if $g(P^N_{j,(i,j)}) < 0$. In contrast, if $g(P^N_{j,(i,j)}) \geq 0$, NOMA becomes useless for user i and user j.

- When $P^o_{i,(i,j)} \leq \frac{q}{2}$ and $q - P_0 \leq P^o_{j,(i,j)}$, let $g_{\text{min}}(x) = \min\{g(x), g_{\mathcal{O}_i}(q/2), g_{\mathcal{O}_j}(\bar{q})\}$, then
$$P^N_{j,(i,j)} = \begin{cases} x, & \text{if } g_{\text{min}}(x) = g(x) \\ q/2, & \text{if } g_{\text{min}}(x) = g_{\mathcal{O}_i}(q/2) \\ \bar{q}, & \text{if } g_{\text{min}}(x) = g_{\mathcal{O}_j}(\bar{q}), \end{cases}$$
where
$$x = \begin{cases} P^o_{j,(i,j)}, & \text{if } \bar{q} < P_j^* < P^o_{j,(i,j)} \\ P_j^*, & \text{if } P^o_{j,(i,j)} < P_j^* < P^o_{i,(i,j)} \\ P^o_{i,(i,j)}, & \text{if } P^o_{i,(i,j)} < P_j^*. \end{cases} \quad (46)$$
- When $P^o_{j,(i,j)} \leq \frac{q}{2} \leq P^o_{i,(i,j)}$ and $q - P_0 \leq P^o_{j,(i,j)}$, let $g_{\text{min}}(x) = \min\{g(x), g_{\mathcal{O}_j}(\bar{q})\}$; then,
$$P^N_{j,(i,j)} = \begin{cases} x, & \text{if } g_{\text{min}}(x) = g(x) \\ \bar{q}, & \text{if } g_{\text{min}}(x) = g_{\mathcal{O}_j}(\bar{q}), \end{cases}$$
where
$$x = \begin{cases} P^o_{j,(i,j)}, & \text{if } \bar{q} < P_j^* < P^o_{j,(i,j)} \\ P_j^*, & \text{if } P^o_{j,(i,j)} < P_j^* < q/2 \\ q/2, & \text{if } q/2 < P_j^*. \end{cases} \quad (47)$$
- When $P^o_{i,(i,j)} \leq \frac{q}{2}$ and $P^o_{j,(i,j)} \leq q - P_0 \leq P^o_{i,(i,j)}$, let $g_{\text{min}}(x) = \min\{g(x), g_{\mathcal{O}_i}(q/2)\}$; then,
$$P^N_{j,(i,j)} = \begin{cases} x, & \text{if } g_{\text{min}}(x) = g(x) \\ q/2, & \text{if } g_{\text{min}}(x) = g_{\mathcal{O}_i}(q/2), \end{cases}$$
where
$$x = \begin{cases} q - P_0, & \text{if } P_j^* < q - P_0 \\ P_j^*, & \text{if } q - P_0 < P_j^* < P^o_{i,(i,j)} \\ P^o_{i,(i,j)}, & \text{if } P^o_{i,(i,j)} < P_j^*. \end{cases} \quad (48)$$
- When $P^o_{j,(i,j)} \leq q - P_0 \leq q/2 \leq P^o_{i,(i,j)}$, then
$$P^N_{j,(i,j)} = \begin{cases} q - P_0, & \text{if } P_j^* < q - P_0 \\ P_j^*, & \text{if } q - P_0 < P_j^* < P^o_{i,(i,j)} \\ P^o_{i,(i,j)}, & \text{if } P^o_{i,(i,j)} < P_j^*. \end{cases} \quad (49)$$
- When $P^o_{i,(i,j)} \leq q - P_0$,
$$P^N_{j,(i,j)} = q/2. \quad (50)$$
- When $\frac{q}{2} \leq P^o_{j,(i,j)}$,
$$P^N_{j,(i,j)} = \bar{q}. \quad (51)$$
- When $\bar{q} \leq P^o_{i,(i,j)} < P^o_{j,(i,j)} < q/2$, let $g_{\text{min}} = \min\{g_{\mathcal{O}_j}(\bar{q}), g_{\mathcal{O}_i}(\frac{q}{2})\}$; then,
$$P^N_{j,(i,j)} = \begin{cases} P^o_{i,(i,j)}, & \text{if } g_{\text{min}} = g_{\mathcal{O}_j}(P^o_{i,(i,j)}) \\ P^o_{j,(i,j)}, & \text{if } g_{\text{min}} = g_{\mathcal{O}_i}(P^o_{j,(i,j)}). \end{cases} \quad (52)$$
- When $\bar{q} \leq P^o_{i,(i,j)}$ and $\frac{q}{2} \leq P^o_{j,(i,j)}$, then
$$P^N_{j,(i,j)} = \bar{q}. \quad (53)$$
- When $P^o_{i,(i,j)} < \bar{q}$ and $P^o_{j,(i,j)} < \frac{q}{2}$, then
$$P^N_{j,(i,j)} = \frac{q}{2}. \quad (54)$$

Proof: The constraints (38) and (39) can be combined together, which is written as

$$\bar{q} \leq P_j \leq q/2 (\leq P_0). \quad (55)$$

Differentiating the objective function $g(P_j)$ by P_j,

$$\frac{\mathrm{d}g}{\mathrm{d}P_j} = \frac{\Gamma_i(\tau Q_i + \tilde{Z}_i)}{N\Gamma_i P_j/2 + 1} - \frac{\Gamma_j(\tau Q_j + \tilde{Z}_j)}{N\Gamma_j P_j/2 + 1}. \quad (56)$$

The local minimizer P_j^* is obtained from $\frac{\mathrm{d}g}{\mathrm{d}P_j} = 0$, as given by (45). The second derivative of the objective function $g(P_j)$ becomes

$$\frac{\mathrm{d}^2 g}{\mathrm{d}P_j^2} = \frac{N\Gamma_i^2\Gamma_j^2(\tau Q_j + \tilde{Z}_j - \tau Q_i - \tilde{Z}_i)^2}{(\Gamma_i - \Gamma_j)^2} \times \Big(\frac{1}{\tau Q_j + \tilde{Z}_j} - \frac{1}{\tau Q_i + \tilde{Z}_i}\Big). \quad (57)$$

Since we already suppose that $1 < \frac{\tau Q_i + \tilde{Z}_i}{\tau Q_j + \tilde{Z}_j} < \frac{\Gamma_j}{\Gamma_i}$, then $P_j^* \geq 0$ and $\frac{\mathrm{d}^2 g}{\mathrm{d}P_j^2} > 0$; consequently, P_j^* is the local minimizer of $g(P_j)$.

Consider the case $P^o_{j,(i,j)} \leq P^o_{i,(i,j)}$ first. The optimal point should be carefully found depending on the relative positions of (55), $\mathcal{O}_i$, and $\mathcal{O}_j$. The objective functions in $\mathcal{O}_j$ and $\mathcal{O}_i$ are $g_{\mathcal{O}_j}(P_j)$ and $g_{\mathcal{O}_i}(P_j)$, respectively. Therefore, three objective functions of $g(P_j)$, $g_{\mathcal{O}_j}(P_j)$, and $g_{\mathcal{O}_i}(P_j)$ are compared to determine the optimal power level depending on the value of P^*_j. For example, consider the case of $q - P_0 \leq P^o_{j,(i,j)}$ and $P^o_{i,(i,j)} \leq \frac{q}{2}$, then $\mathcal{O}_j = [q - P_0, P^o_{j,(i,j)}]$ and $\mathcal{O}_i = [P^o_{i,(i,j)}, \frac{q}{2}]$. In addition, $g_{\mathcal{O}_j}(\bar{q})$ and $g_{\mathcal{O}_i}(q/2)$ are minimizers in $\mathcal{O}_j$ and $\mathcal{O}_i$, respectively.

However, the minimizer in $[P^o_{j,(i,j)}, P^o_{i,(i,j)}]$ depends on P^*_j. If $P^o_{j,(i,j)} \leq P^*_j \leq P^o_{i,(i,j)}$, P^*_j minimizes $g(P_j)$ obviously. On the other hand, if $P^*_j < P^o_{j,(i,j)}$, $P^o_{j,(i,j)}$ becomes the minimizer of $g(P_j)$ in $[P^o_{j,(i,j)}, P^o_{i,(i,j)}]$. Similarly, if $P^o_{i,(i,j)} < P^*_j$, $P^o_{i,(i,j)}$ is the minimizer of $g(P_j)$ in $[P^o_{j,(i,j)}, P^o_{i,(i,j)}]$. Therefore, the optimal objective value in $[P^o_{j,(i,j)}, P^o_{i,(i,j)}]$ is given by $g(x)$, where

$$x = \begin{cases} P^o_{j,(i,j)}, & \text{if } P^*_j < P^o_{j,(i,j)} \\ P^*_j, & \text{if } P^o_{j,(i,j)} \leq P^*_j \leq P^o_{i,(i,j)} \\ P^o_{i,(i,j)}, & \text{if } P^o_{i,(i,j)} < P^*_j. \end{cases} \tag{58}$$

Finally, the globally optimal objective function can be obtained by taking the minimal one among $g(x)$, $g_{\mathcal{O}_j}(\bar{q})$, and $g_{\mathcal{O}_i}(q/2)$, as shown in (46). In a similar way, the solution of (47)–(50) also can be derived depending on the relative positions of (55), $\mathcal{O}_i$, and $\mathcal{O}_j$.

Next, at least one of the users experiences the link outage in the case of $P^o_{i,(i,j)} \leq P^o_{j,(i,j)}$. We can define the interval of $\mathcal{O}_{(i,j)} = [P^o_{i,(i,j)}, P^o_{j,(i,j)}]$ as the outage region of both users. If $P^N_{j,(i,j)} \in \mathcal{O}_{(i,j)}$, both links are in outage so NOMA becomes meaningless. Therefore, we just need to compare the objective function values of $g_{\mathcal{O}_i}$ and $g_{\mathcal{O}_j}$. For example, when $\bar{q} \leq P^o_{i,(i,j)}$, the minimal objective function value in $\mathcal{O}_j$ is $g_{\mathcal{O}_j}(\bar{q})$. On the other hand, when $P^o_{j,(i,j)} \leq \frac{q}{2}$, the minimal objective function value in $\mathcal{O}_i$ is $g_{\mathcal{O}_i}(\frac{q}{2})$. Thus, the solution to (52) is obtained by choosing the minimum of $g_{\mathcal{O}_j}(\bar{q})$ and $g_{\mathcal{O}_i}(\frac{q}{2})$. In addition, $P^o_{i,(i,j)} < \bar{q}$ and $\frac{q}{2} < P^o_{j,(i,j)}$ mean that $P^N_{j,(i,j)}$ cannot be in $\mathcal{O}_j$ and $\mathcal{O}_i$, respectively, so (53) and (54) can be obtained directly. □

Remark: Even though two users are paired for NOMA, no transmit power could be allocated to one of the users. This case indicates that the resource of one of the users is taken by another one. For example, when an outage occurs at a certain link, the resource of this link is preferentially utilized by another link for resource efficiency. Thus, we can see that finding the optimal power in this model enables link scheduling as well as the flexible use of system resources.

The second subproblem for resolving the power allocation problem of (35)–(36) involves finding the optimal auxiliary variable of q to minimize the optimization metric of (35) as follows:

$$q^N = \arg\min_{q} h(q) \tag{59}$$

$$\text{s.t. } 0 \leq q \leq 2P_0, \tag{60}$$

where

$$h(q) = V \cdot q - (\tau Q_i + \tilde{Z}_i)\tilde{R}_i(q) - (\tau Q_j + \tilde{Z}_j)\tilde{R}_j. \tag{61}$$

Here, $R_i(q) = \frac{2\Phi\mathcal{B}}{N}\log_2\left(1 + \frac{N\Gamma_i q/2+1}{N\Gamma_i P^N_{j,(i,j)}/2+1}\right)$ and $R_j = \log_2(1 + N\Gamma_j P^N_{j,(i,j)}/2)$.

Differentiating $h(q)$ by q,

$$\frac{\mathrm{d}h}{\mathrm{d}q} = V - \frac{\tau Q_i + Z_i}{\ln 2} \cdot \frac{\Phi\mathcal{B}\Gamma_i}{N\Gamma_i q/2 + 1}, \tag{62}$$

and the local minimizer q^* can be obtained from $\frac{\mathrm{d}h}{\mathrm{d}q} = 0$ by

$$q^* = \frac{2}{N\Gamma_i}\Big(\frac{\Phi\mathcal{B}\Gamma_i}{V\ln 2}(\tau Q_i + Z_i) - 1\Big). \tag{63}$$

Since

$$\frac{\mathrm{d}^2\, h}{\mathrm{d}q^2} = \frac{\tau Q_i + Z_i}{\ln 2} \cdot \frac{\mathcal{B}\Gamma_i}{(N\Gamma_i q/2 + 1)^2} \cdot \frac{N\Gamma_i}{2} > 0, \tag{64}$$

q^* is the global minimizer of $h(q)$. Considering the constraint (60), the optimal q is obtained by

$$q^N = \begin{cases} 0, & \text{if } q^* < 0 \\ q^*, & \text{if } 0 \leq q^* \leq 2P_0 \\ 2P_0, & \text{if } 2P_0 < q^*. \end{cases} \tag{65}$$

Herein, $q^N = 0$ makes NOMA useless, because both users are in outage.

Thus, the non-convex optimization problem of (35)–(36) for finding the optimal power allocation for NOMA users can be solved by dealing with the two convex subproblems of (59)–(60) and (37)–(39) sequentially. The transmitter can first optimize the transmit power consumption for given user pair, i.e., q^N, based on the current CSI and QSI. Then, power levels allocated to the NOMA users, i.e., $P^N_{i,(i,j)}$ and $P^N_{j,(i,j)}$, can be achieved by Theorem 2.

V. Matching Algorithm for NOMA User Pairing

Since the optimal power allocation rule is derived when the pair of NOMA users is already determined, there remains the problem of which users are better to be paired for NOMA signaling. User pairing can be interpreted as a matching problem. We now define the matching Ψ which indicates user pairings for NOMA signaling, and we change the notation from $\Psi_m = n$ to $\Psi(m) = n$ in this section for utilizing the matching theory.

Definition 1: A matching Ψ is defined by (66)–(68) as follows:

$$\Psi(u_i) \in \mathcal{U},\ \forall u_i \in \mathcal{U} \tag{66}$$

$$|\Psi(u_i)| = 1 \tag{67}$$

$$\Psi(u_i) = u_j \iff \Psi(u_j) = u_i. \tag{68}$$

Specifically, $\Psi(u_i)$ indicates the user paired with user u_i and both users are in the same user set $\mathcal{U}$ consisting of N users; therefore, (66) is satisfied. $\Psi(u_i) = u_i$ means that OMA is used for u_i, and $\Psi(u_i) = u_j$ for $i \neq j$ indicates that u_i and u_j are paired for NOMA. Since we considered the two-user NOMA model, (67) is given and Ψ becomes the one-to-one matching. When u_i and u_j are paired, both $\Psi(u_i) = u_j$ and $\Psi(u_j) = u_i$ are satisfied as shown in (68); therefore, $\Psi = \Psi^{-1}$. In previous sections, user pairing is denoted by $\Psi_n = m$ and $\Psi_m = n$ which means users m and n are paired for NOMA.

The matching Ψ is constructed according to the preference lists of users. Denote the preference list of u_i by $\mathcal{P}_i$ for all $u_i \in \mathcal{U}$. When $\mathcal{M}^N_{i,(i,j)} < \mathcal{M}^N_{i,(i,k)}$, u_i prefers u_j to u_k to be paired with. Herein, $\mathcal{M}^N_{i,(i,i)} = \mathcal{M}^O_i$. In addition, we only allow u_j to be included in $\mathcal{P}_i$ when $\mathcal{M}^N_{i,(i,j)} \leq 0$. The reason is that the condition $\mathcal{M}^N_{i,(i,j)} = 0$ for any u_j is obtained when $P_i = 0$. Before constructing Ψ, each optimization metric $\mathcal{M}^N_{i,(i,j)}$ for all $u_i, u_j \in \mathcal{U}$ can be obtained by solving the problems of (26)–(27) and (35)–(36) for $i = j$ and $i \neq j$, respectively. Given Ψ, the optimal power allocation rule can be obtained according to Section IV-B and is denoted by $\mathbf{P}^*(\mathbf{\Psi})$. Then, the total optimization metric of (23) can be computed as $\mathcal{M}(\mathbf{P}^*(\mathbf{\Psi}), \mathbf{\Psi}) = \sum_{i \in \mathcal{N}} \mathcal{M}^N_{i,(i,\Psi(u_i))}$.

Since there is too much complexity involved with computing and comparing the optimization metrics for all possible pairing combinations, we simply focus on seeking the stable matching by using the deferred acceptance (DA) procedure [50]. Each user sends the matching request to the most preferred user, and the user who receives the request can accept or reject pairing with the sender. The user pairing algorithm is shown in Algorithm 1, and the details of the matching request and decision for the received request are expressed in Algorithm 2.

Algorithm 1 User Pairing Algorithm for NOMA Transmissions

```
Initialize $\Psi(u) \leftarrow u,\ \forall u \in \mathcal{U}$.
for $\forall u_i \in \mathcal{U}$ do
    $\mathcal{F} \leftarrow \phi$
    while true do
        Find $u_j \leftarrow \arg\min_{u \in \mathcal{P}_i \setminus \mathcal{F}} \mathcal{M}^N_{i,(i,j)}$
        if $j == \Psi(u_i)$ then
            break;
        end if
        $\mathcal{F} \leftarrow \mathcal{F} \cup \{u_j\}$
        $\Psi' \leftarrow$ MatchRequest$(u_i, u_j, \Psi, \phi)$
        if $\mathcal{M}(\mathbf{P}(\mathbf{\Psi}), \mathbf{\Psi}) > \mathcal{M}(\mathbf{P}(\mathbf{\Psi'}), \mathbf{\Psi'})$ then
            $\Psi \leftarrow \Psi'$
            break;
        end if
    end while
end for
```

For example, suppose that u_i sends the matching request to u_j in the matching Ψ. Let Ψ' be the *optimal* matching when u_j accepts the request from u_i. Then, u_j decides whether to accept or reject the request from u_i by comparing $\mathcal{M}(\mathbf{P}^*(\mathbf{\Psi}), \mathbf{\Psi})$ and $\mathcal{M}(\mathbf{P}^*(\mathbf{\Psi'}), \mathbf{\Psi'})$. $\mathcal{M}(\mathbf{P}^*(\mathbf{\Psi}), \mathbf{\Psi})$ is already obtained with the current matching Ψ, and we need to compute $\mathcal{M}(\mathbf{P}^*(\mathbf{\Psi'}), \mathbf{\Psi'})$. When $\Psi(u_i) = u_i$ and $\Psi(u_j) = u_j$, the matching request is simply accepted when $\mathcal{M}^N_{i,(i,j)}(P^N_{i,(i,j)}) + \mathcal{M}^N_{j,(i,j)}(P^N_{j,(i,j)}) < \mathcal{M}^O_i(P^O_i) + \mathcal{M}^O_j(P^O_j)$. Then, the optimal matching becomes $\Psi' = \Psi \setminus \{(u_i, u_i), (u_j, u_j)\} \cup \{(u_i, u_j), (u_j, u_i)\}$.

Algorithm 2 Matching Request Algorithm

```
Input: $u_i$, $u_j$, $\Psi'$, and $\mathcal{E}$.
Output: $\Psi'$
$m \leftarrow \Psi'(u_i)$ and $p \leftarrow \Psi'(u_j)$
$\Psi(u_i) \leftarrow u_j$ and $\Psi(u_j) \leftarrow u_i$
if $i \neq m$ then $\Psi'(u_m) \leftarrow u_m$
end if
if $j \neq p$ then $\Psi'(u_p) \leftarrow u_p$
end if
$\mathcal{E}_m \leftarrow \mathcal{E} \cup \{u_i, u_j\}$
if $i \neq m$ then
    Find $n \leftarrow \arg\min_{u_n \in \mathcal{P}_m \setminus \mathcal{E}_m} \mathcal{M}^N_{m,(m,n)}$
    if $n == m$ then $\Psi'(u_n) \leftarrow u_n$
    else if $n == p$ then $\Psi'(u_n) \leftarrow u_p$ and $\Psi'(u_p) \leftarrow u_n$
    else $\Psi' \leftarrow$ MatchRequest$(u_m, u_n, \Psi', \mathcal{E}_m)$
    end if
end if
$\mathcal{E}_p \leftarrow \mathcal{E} \cup \{u_i, u_j\}$
if $j \neq p$ && $\Psi'(u_p) == u_p$ then
    Find $q \leftarrow \arg\min_{u_q \in \mathcal{P}_p \setminus \mathcal{E}_p} \mathcal{M}^N_{p,(p,q)}$
    if $q == p$ then $\Psi'(u_q) \leftarrow u_q$
    else $\Psi' \leftarrow$ MatchRequest$(u_p, u_q, \Psi', \mathcal{E}_p)$
    end if
end if
return $\Psi'$
```

However, when $\Psi(u_i) \neq u_i$ and/or $\Psi(u_j) \neq u_j$, if u_j accepts the matching request from u_i, $\Psi(u_i)$ and/or $\Psi(u_j)$ should find another pair to construct the optimal matching Ψ'. According to Algorithm 2, $\Psi(u_i)$ and $\Psi(u_j)$ send the matching request to their most preferred users, except for u_i and u_j. If the most preferred users of $\Psi(u_i)$ and $\Psi(u_j)$ are themselves respectively, then $\Psi' = \Psi \setminus \{(u_i, \Psi(u_i)), (u_j, \Psi(u_j))\} \cup \{(u_i, u_j), (u_j, u_i), (\Psi(u_i), \Psi(u_i)), (\Psi(u_j), \Psi(u_j))\}$. If not, Algorithm 2 should be recursively performed to construct Ψ' until all users are matched. Finally, compute $\mathcal{M}(\mathbf{P}^*(\mathbf{\Psi'}), \mathbf{\Psi'})$ and compare it with $\mathcal{M}(\mathbf{P}^*(\mathbf{\Psi}), \mathbf{\Psi})$. If $\mathcal{M}(\mathbf{P}^*(\mathbf{\Psi}), \mathbf{\Psi}) > \mathcal{M}(\mathbf{P}^*(\mathbf{\Psi'}), \mathbf{\Psi'})$, the match request from u_i to u_j is accepted and Ψ is updated by Ψ', as shown in Algorithm 1.

The optimal user pairing can be obtained by searching over all possible combinations of user pairings to find $\mathbf{\Psi}$ that maximizes $\mathcal{M}(\mathbf{P}^*(\mathbf{\Psi}), \mathbf{\Psi})$. Suppose that L pairs are allowed for NOMA. Then, the transmitter needs to exhaustively search $\prod_{l=1}^{L} \binom{N-2(l-1)}{2}$ combinations for the optimal user pairing, and the time complexity is approximately $O(N^{2L})$. In the proposed algorithm, the worst case is that no pair is generated for $u_1, \cdots, u_{N-L}$, and then a new pair is matched every time for the last L users. This requires $\sum_{l=1}^{L} N - 2(l-1) + N(N-L)$ comparison steps, and the time complexity is $O(N^2)$. Thus, the complexity of the proposed matching algorithm is much less than that of the optimal user pairing. Note that the complexity gain of the proposed algorithm grows as N and L increase.

VI. Performance Evaluation

Our data-intensive simulations for the performance evaluation are based on the cellular model with a radius $R = 50$. There exist $N = 40$ users, and all users are uniformly placed in the whole cellular region. Assume that $\mathcal{N}_s = \mathcal{N}$, $\rho = \rho_n$, and

TABLE I

SYSTEM PARAMETERS [11], [12]

E2E delay bound (D_{max})	1 ms
Frame duration (T_f)	0.1 ms
DL phase duration (T_D)	0.05 ms
Maximal queueing delay (D^q_{max})	0.9 ms
Packet size (u)	160 bits
Cell radius (R)	50 m
Path loss model	35.3+37.6(d_k)
Single-sided noise spectral density (N_0)	-173 dBm/Hz
Power budget for each user (P_0)	3 W
User number (N)	40
Bandwidth (BW)	20 MHz
V	5×10^5
λ_{min}	5 packets
λ_{max}	10 packets
ρ	7 Mb
η	8.5 Mb

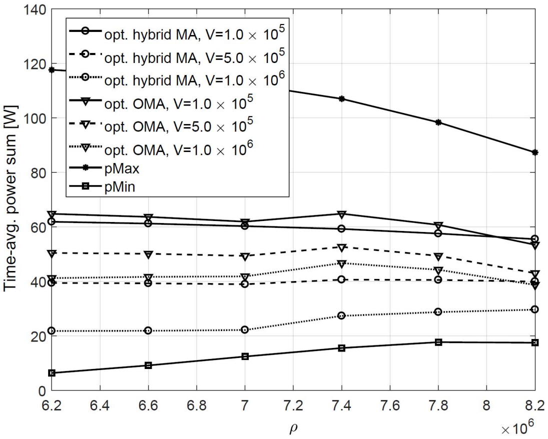

Fig. 3. Time-avg. transmit power sum vs. ρ.

$\eta = \eta_n$ for all $n \in \mathcal{N}$. All parameters are listed in Table I, and these are used unless otherwise noted. A short frame structure designed for low-latency communications in 5G networks [12] is used for simulation. Note that the maximum queueing delay bound is $D^q_{max} = 0.9$ ms, and we will show that the proposed algorithm achieves this queueing delay constraint.

To verify the advantages of the proposed algorithm, this paper compares the proposed one with the following other schemes:

- "pMax": The transmitter always consumes the maximum power budget for all of the N users, except when a link outage occurs. NOMA and user pairing are not considered.
- "pMin": The transmitter always consumes the minimum power enough to avoid the link outage. If the required power for avoiding the outage is greater than the power budget, the link remains in outage. NOMA and user pairing are not considered.
- "opt. OMA": The power allocation is based on the proposed optimization framework (7)-(10) but NOMA and user pairing are not considered.

To emphasize the difference from the above comparison schemes, we will call the proposed scheme "opt. hybrid MA".

Figs. 3 and 4 show plots of the time-average transmit power consumption for N users and the expected maximum queueing delay among N users versus ρ, i.e., the outage threshold, respectively. In addition, the time-average data rates and delay-constrained transmission rates with different values of ρ are shown in Figs. 5 and 6, respectively. The delay-constrained transmission rate means the probability that a data packet arriving at the queue can be transmitted within

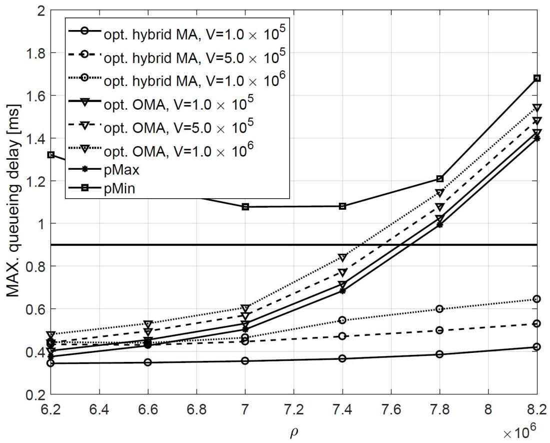

Fig. 4. Maximal queueing delay vs. ρ.

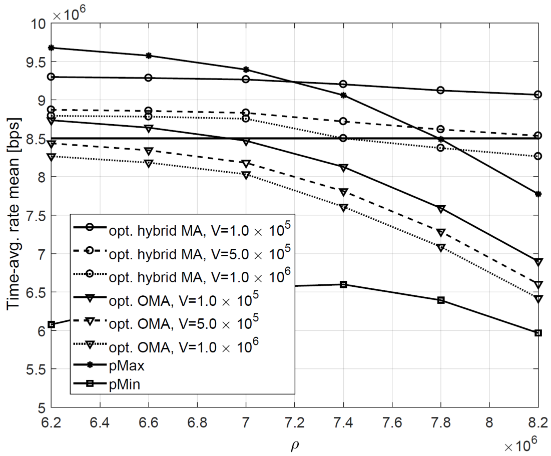

Fig. 5. Time-avg. data rate vs. ρ.

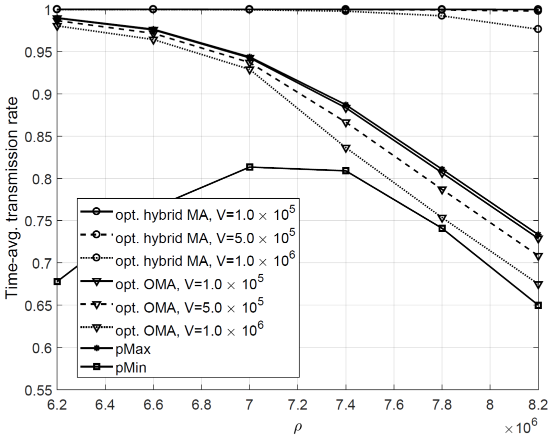

Fig. 6. Delay-constrained transmission rate vs. ρ.

the time D^q_{max}. The link outage occurs more frequently as ρ increases. Therefore, increasing ρ basically makes the power consumption and queueing delay grow, and the time-average data rate and the delay-constrained transmission rate decrease.

However, some peculiar trends can be observed in Figs. 3–6. First, some cases show decreasing power consumption as ρ increases, and this results from frequent link deactivations due to link outage occurrences. Second, the performance trends of pMin are not monotonic in Figs. 4–6. The reason for this is that the activated link of pMin always provides the data rate of ρ because pMin consumes the minimum power to avoid the outage. Therefore, the time-average data rate of pMin increases with small ρ values, but it decreases with ρ, when ρ is large because of the frequent outage occurrences.

Among comparison techniques, we can notice that the proposed opt. hybrid MA provides high power-efficiency while guaranteeing the low queueing delay and the sufficient time-average data rate. The opt. hybrid MA with $V = 1.0\times 10^6$ consumes almost the same power as pMin in Fig. 3; moreover, the queueing delay of opt. hybrid MA with $V = 1.0 \times 10^5$ is the shortest among the comparison schemes in Fig. 4. Especially when ρ is large, all of the other schemes require maximum queueing delays larger than D^q_{max}, but opt. hybrid MA does not. Therefore, opt. hybrid MA also shows the best delay-constrained transmission rates given in Fig. 6. In addition, opt. hybrid MA satisfies the QoS constraint as shown in Fig. 5.

We can also see the advantages of NOMA over OMA by comparing opt. hybrid MA with opt. OMA. NOMA is well-known to improve throughput compared to OMA, with the same power consumption. We can see that opt. hybrid MA gives better data rates with smaller power consumption than opt. OMA in Figs. 3 and 5. The proposed opt. hybrid MA guarantees the sufficiently large time-average data rate and flexibly utilizes NOMA advantages over OMA in terms of both data rate and power-efficiency. Further, lower queueing delays and higher delay-constrained transmission rates are achieved by using NOMA with appropriate user pairings and power allocations.

The effects of the system parameter V are also shown in Figs. 3–6. As we explained earlier, V is a weight factor for the term representing the transmit power in (23). As V becomes larger, opt. hybrid MA and opt. OMA further pursue power-efficiency rather than reducing queue backlogs, i.e., $Q_n(t)$ and $Z_n(t)$, for all $n \in \mathcal{N}$. Therefore, the time-average transmit power decreases with V. On the other hand, $Q_n(t)$ and $Z_n(t)$ grow as V increases; therefore, the queueing delay increases. To that effect, the delay-constrained transmission rate decreases and the time-average data rate decreases in opt. hybrid MA and opt. OMA. Thus, the tradeoff between power consumption and queueing delay can be controlled by adjusting the system parameter V, depending on the stochastic networks and QoS requirements.

Figs. 7–10 show plots of power consumption, queueing delay, time-average data rate, and delay-constrained transmission rate versus η, respectively. The QoS constraint η only affects the performances of opt. hybrid MA and opt. OMA because pMax and pMin do not consider the QoS constraint.

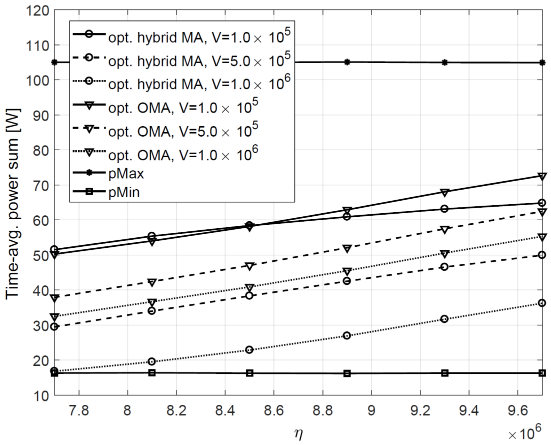

Fig. 7. Time-avg. transmit power sum vs. η.

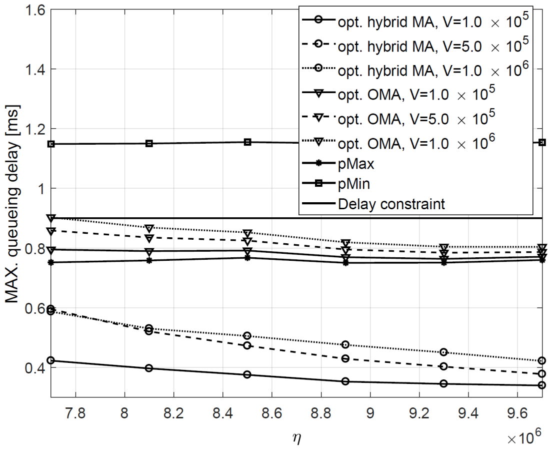

Fig. 8. Maximal queueing delay vs. η.

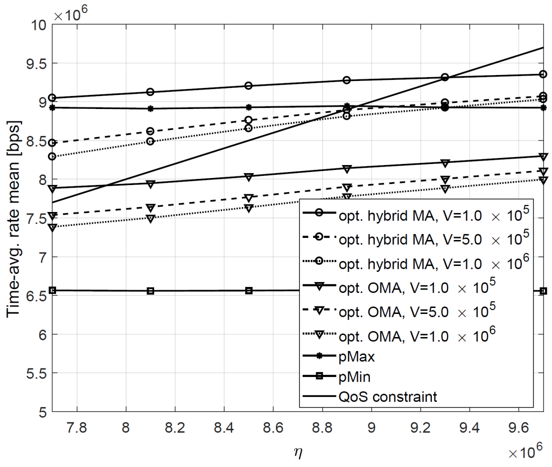

Fig. 9. Time-avg. data rate vs. η.

As η grows, transmit powers of opt. hybrid MA and opt. OMA obviously increase to guarantee η as much as possible; therefore, the time-average data rates of opt. hybrid MA and opt. OMA increase. Since the larger power causes more departures, their queueing delays decrease as well. However, it becomes

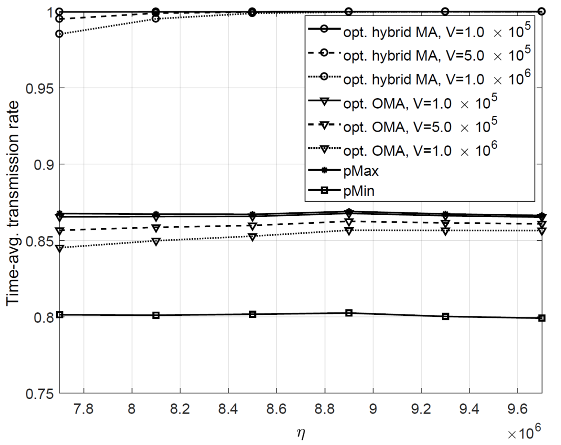

Fig. 10. Delay-constrained transmission rate vs. η.

too difficult to satisfy QoS requirements when η is large, as shown in Fig. 9. The encouraging point is that opt. hybrid MA can provide higher time-average data rates even with less power consumption than pMax. In addition, by comparing opt. hybrid MA with opt. OMA, we can see that NOMA advantages over OMA still remain for different values of η. Similar to performance changes with V in Figs. 3–6, the time-average transmit power decreases as V grows, in contrast, the queueing delay increases, as shown in Figs. 7 and 8, respectively.

VII. Concluding Remarks

This paper presents joint optimization framework for power allocation and user pairing in the hybrid MA system. The optimization framework pursues both power-efficiency and low queueing delay while achieving sufficient time-average data rates. User pairings for NOMA signaling are performed based on the matching theory, and the closed-form optimal power allocations for OMA and NOMA users with a given policy of user pairing are derived. The proposed algorithm dynamically conducts user pairing steps for NOMA with optimal power allocations to adjust backlogs in transmitter queues. Based on the short frame structure, simulation results show that the proposed algorithm enables one to satisfy the delay constraint, while guaranteeing high power efficiency and sufficient time-average data rates. The proposed dynamic power control and user pairing algorithm smooths out the tradeoff between power consumption and queueing delay.

References

[1] J. G. Andrews *et al.*, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[2] L. Wei, R. Q. Hu, Y. Qian, and G. Wu, "Enable device-to-device communications underlaying cellular networks: Challenges and research aspects," *IEEE Commun. Mag.*, vol. 52, no. 6, pp. 90–96, Jun. 2014.

[3] S. Buzzi, C.-L. I, T. E. Klein, H. V. Poor, C. Yang, and A. Zappone, "A survey of energy-efficient techniques for 5G networks and challenges ahead," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 697–709, Apr. 2016.

[4] R. Q. Hu and Y. Qian, "An energy efficient and spectrum efficient wireless heterogeneous network framework for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 94–101, May 2014.

[5] A. Aijaz and A. H. Aghvami, "Cognitive machine-to-machine communications for Internet-of-Things: A protocol stack perspective," *IEEE Internet Things J.*, vol. 2, no. 2, pp. 103–112, Apr. 2015.

[6] J. J. Nielsen, R. Liu, and P. Popovski, "Ultra-reliable low latency communication using interface diversity," *IEEE Trans. Commun.*, vol. 66, no. 3, pp. 1322–1334, Mar. 2018.

[7] Z. Chen, N. Pappas, M. Kountouris, and V. Angelakis, "Throughput analysis of smart objects with delay constraints," in *Proc. IEEE 17th Int. Symp. World Wireless, Mobile Multimedia Netw.*, Coimbra, Portugal, Jun. 2016, pp. 1–6.

[8] S. Zuo, I.-H. Hou, T. Liu, A. Swami, and P. Basu, "Joint rate control and scheduling for real-time wireless networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4562–4570, Jul. 2017.

[9] C. She, C. Yang, and T. Q. S. Quek, "Radio resource management for ultra-reliable and low-latency communications," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 72–78, Jun. 2017.

[10] C. She, C. Yang, and T. Q. S. Quek, "Joint uplink and downlink resource configuration for ultra-reliable and low-latency communications," *IEEE Trans. Commun.*, vol. 66, no. 5, pp. 2266–2280, May 2018.

[11] C. She and C. Yang, "Ensuring the quality-of-service of tactile Internet," in *Proc. IEEE 83rd Veh. Technol. Conf. (VTC Spring)*, Nanjing, China, May 2016, pp. 1–5.

[12] P. Kela *et al.*, "A novel radio frame structure for 5G dense outdoor radio access networks," in *Proc. IEEE 81st Veh. Technol. Conf. (VTC Spring)*, Glasgow, U.K., May 2015, pp. 1–6.

[13] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, Jul. 2003.

[14] C. She, C. Yang, and T. Q. S. Quek, "Cross-layer transmission design for tactile Internet," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Washington, DC, USA, Dec. 2016, pp. 1–6.

[15] C. She, C. Yang, and T. Q. S. Quek, "Cross-layer optimization for ultra-reliable and low-latency radio access networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 127–141, Jan. 2018.

[16] D. Bertsekas and G. R. Gallager, *Data Networks*, 2nd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 1992.

[17] M. J. Neely and S. Supittayapornpong, "Dynamic Markov decision policies for delay constrained wireless scheduling," *IEEE Trans. Autom. Control*, vol. 58, no. 8, pp. 1948–1961, Aug. 2013.

[18] M. Choi, J. Kim, and J. Moon, "Adaptive detector selection for queue-stable word error rate minimization in connected vehicle receiver design," *IEEE Trans. Veh. Tech.*, vol. 67, no. 4, pp. 3635–3639, Apr. 2018.

[19] M. Choi *et al.*, "Wireless video caching and dynamic streaming under differentiated quality requirements," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1245–1257, Jun. 2018.

[20] P.-C. Hsieh, I.-H. Hou, and X. Liu, "Delay-optimal scheduling for queueing systems with switching overhead," 2017, *arXiv:1701.03831*. [Online]. Available: https://arxiv.org/abs/1701.03831

[21] R. A. Berry and R. G. Gallager, "Communication over fading channels with delay constraints," *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1135–1149, May 2002.

[22] R. A. Berry, "Optimal power-delay tradeoffs in fading channels—Small-delay asymptotics," *IEEE Trans. Inf. Theory*, vol. 59, no. 6, pp. 3939–3952, Jun. 2013.

[23] C. Sun, C. She, C. Yang, T. Q. S. Quek, Y. Li, and B. Vucetic, "Optimizing resource allocation in the short blocklength regime for ultra-reliable and low-latency communications," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 402–415, Jan. 2019.

[24] A. Fu, E. Modiano, and J. Tsitsiklis, "Optimal energy allocation for delay-constrained data transmission over a time-varying channel," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, San Francisco, CA, USA, vol. 2, 2003, pp. 1095–1105.

[25] J. Lee and N. Jindal, "Energy-efficient scheduling of delay constrained traffic over fading channels," *IEEE Trans. Wireless Commun.*, vol. 8, no. 4, pp. 1866–1875, Apr. 2009.

[26] I. Bettesh and S. S. Shamai (Shitz), "Optimal power and rate control for minimal average delay: The single-user case," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 4115–4141, Sep. 2006.

[27] M. Goyal, A. Kumar, and V. Sharma, "Power constrained and delay optimal policies for scheduling transmission over a fading channel," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, vol. 1, Mar./Apr. 2003, pp. 311–320.

[28] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation over wireless links," *IEEE Trans. Wireless Commun.*, vol. 6, no. 8, pp. 3058–3068, Aug. 2007.

[29] M. J. Neely, "Optimal energy and delay tradeoffs for multiuser wireless downlinks," *IEEE Trans. Inf. Theory*, vol. 53, no. 9, pp. 3095–3113, Sep. 2007.

[30] S. Supittayapornpong and M. J. Neely, "Achieving utility-delay-reliability tradeoff in stochastic network optimization with finite buffers," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Hong Kong, Apr./May 2015, pp. 1427–1435.

[31] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. Bhargava, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181–2195, Oct. 2017.

[32] T. David and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[33] M. Shirvanimoghaddam *et al.*, "Massive non-orthogonal multiple access for cellular IoT: Potentials and limitations," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 55–61, Sep. 2017.

[34] X. Sun, S. Yan, N. Yang, Z. Ding, C. Shen, and Z. Zhong, "Short-packet downlink transmission with non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4550–4564, Jul. 2018.

[35] Y. Yu, H. Chen, Y. Li, Z. Ding, and B. Vucetic, "On the performance of non-orthogonal multiple access in short-packet communications," *IEEE Commun. Lett.*, vol. 22, no. 3, pp. 590–593, Mar. 2018.

[36] Z. Ding, M. Peng, and H. V. Poor, "Cooperative non-orthogonal multiple access in 5G systems," *IEEE Commun. Lett.*, vol. 19, no. 8, pp. 1462–1465, Aug. 2015.

[37] M. Choi, D.-J. Han, and J. Moon, "Bi-directional cooperative NOMA without full CSIT," *IEEE Trans. Wireless Commun.*, vol. 17, no. 11, pp. 7515–7527, Nov. 2018.

[38] *Study on Downlink Multiuser Superposition Transmission for LTE*, document TR 36.859, 3GPP, Jan. 2016.

[39] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.

[40] W. Liang, Z. Ding, Y. Li, and L. Song, "User pairing for downlink non-orthogonal multiple access networks using matching algorithm," *IEEE Trans. Commun.*, vol. 65, no. 12, pp. 5319–5332, Dec. 2017.

[41] J.-M. Kang and I.-M. Kim, "Optimal user grouping for downlink NOMA," *IEEE Wireless Commun. Lett.*, vol. 7, no. 5, pp. 724–727, Oct. 2018.

[42] B. Di, L. Song, and Y. Li, "Sub-channel assignment, power allocation, and user scheduling for non-orthogonal multiple access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7686–7698, Nov. 2016.

[43] J. Zhu, J. Wang, Y. Huang, S. He, X. You, and L. Yang, "On optimal power allocation for downlink non-orthogonal multiple access systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2744–2757, Dec. 2017.

[44] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 65, no. 3, pp. 1077–1091, Mar. 2017.

[45] F. Fang, H. Zhang, J. Cheng, and V. C. M. Leung, "Energy-efficient resource allocation for downlink non-orthogonal multiple access network," *IEEE Trans. Commun.*, vol. 64, no. 9, pp. 3722–3732, Sep. 2016.

[46] W. Bao, H. Chen, Y. Li, and B. Vucetic, "Joint rate control and power allocation for non-orthogonal multiple access systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2798–2811, Dec. 2017.

[47] M. J. Neely, "Stochastic network optimization with application to communication and queueing systems," *Synth. Lectures Commun. Netw.*, vol. 3, no. 1, pp. 1–211, 2010.

[48] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[49] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4232–4265, Jun. 2014.

[50] A. Roth and M. A. Sotomayor, *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*. Cambridge U.K.: Cambridge, Univ. Press, 1992.

Minseok Choi received the B.S., M.S., and Ph.D. degrees from the School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2011, 2013, and 2018, respectively. He is currently a Post-Doctoral Researcher of electrical and computer engineering with the University of Southern California (USC), Los Angeles, CA, USA, and a Post-Doctoral Associate at the School of Software, Chung-Ang University, Seoul, South Korea. His research interests include wireless caching networks, stochastic network optimization, non-orthogonal multiple access, and 5G networks.

Joongheon Kim (M'06–SM'18) received the B.S. and M.S. degrees in computer science and engineering from Korea University, Seoul, South Korea, in 2004 and 2006, respectively, and the Ph.D. degree in computer science from the University of Southern California (USC), Los Angeles, CA, USA, in 2014. He has been an Assistant Professor with Korea University, since 2019. Before joining Korea University as a Faculty Member, he was with Chung-Ang University, Seoul, Korea, (2016–2019), Intel Corporation, Santa Clara, CA, USA, (2013–2016), InterDigital, San Diego, CA, USA, in 2012, and LG Electronics Seoul, Korea, from 2006 to 2009. He is a member of the IEEE Communications Society. He was a recipient of the Annenberg Graduate Fellowship with his Ph.D. admission from USC (2009).

Jaekyun Moon (F'05) received the Ph.D. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA. From 1990 to 2009, he was a faculty member at the Department of Electrical and Computer Engineering, University of Minnesota, Twin Cities, MN, USA. He consulted as a Chief Scientist at DSPG, Inc., from 2004 to 2007. He was also the Chief Technology Officer at Link-A-Media Devices Corporation. He is currently a Professor and the Head of the School of Electrical Engineering, KAIST. His research interests include distributed and decentralized storage, communication, and machine intelligence. He was a recipient of the IBM Faculty Development Awards and the IBM Partnership Awards, the National Storage Industry Consortium Technical Achievement Award for the invention of the maximum transition run code, a widely used error-control/modulation code in commercial storage systems, and the McKnight Land-Grant Professorship from the University of Minnesota. He has served as the Program Chair for the 1997 IEEE Magnetic Recording Conference. He was also the Chair of the Signal Processing for Storage Technical Committee of the IEEE Communications Society. He has served as a Guest Editor for the 2001 IEEE Journal on Selected Areas in Communications Issue on Signal Processing for High-Density Recording. He also served as an Editor for the IEEE Transactions on Magnetics in the area of signal processing and coding from 2001 to 2006.