Energy-efficient dual-hop IoT communications network with delay-outage constraints

***Abstract*—This work considers a dual-hop Internet of Thing (IoT) communications network where sensor nodes transmit data to a gateway either directly or via other nodes using dual-hop communications. Each node employs separate transmission buffers to store its own sensing data generated with constant rates, and data received from other nodes. We impose end-to-end delay constraints in terms of the maximum acceptable delay-outage probabilities. We investigate energy-efficient adaptive resource allocation problems (i.e., joint link scheduling, rate, and power allocation) to support minimum data rates of the nodes. A novel approach is proposed exploiting asymptotic delay analysis to first determine the achieved delay exponents (exponential decay rates) of the queue length tail distributions to satisfy the delay-outage constraints. Next, the relation between the delay exponents and resource allocation variables are derived. Last, the solutions to the resulting constrained optimization problems are obtained using Lagrangian approach and convex optimization. Online resource allocation algorithms are developed when the fading statistics are *a-priori* unknown. Illustrative examples are presented to demonstrate the effects of the rate requirements and delay constraint stringency on the power consumption and routing configuration.**

***Index Terms*—Resource allocation, dual-hop communications, delay-outage constraints, wireless sensor network, IoT.**

I. Introduction

The Internet of Things (IoT) allows billions of smart devices to be connected. The devices can be sensors/actuators, able to operate and transmit/receive data to/from other systems without or with minimal human intervention [1]. The rapid IoT development has lead to numerous applications being implemented such as smart factories and manufacturing, smart energy grids, and smart transportation systems, (i.e., the new Industrial Internet of Things (IIoT) paradigm [2]– [4]). IIoT benefits include improved productivity, analytics, and the transformation of the workplace having the potential to generate $15 trillion of global GDP by 2030 [5]. Nevertheless, many research issues still remain open in IIoT [2]– [4]. In particular, as the number of connected devices/sensors is growing exponentially, the ever-increasing energy consumption is one key concern and limitation for the widespread IoT deployment. Hence, reducing energy consumption and improving energy efficiency have become important design challenges in IIoT [6], [7].

Focusing on developing novel energy-efficient radio resource management and transmissions for IIoT applications,[1] the current work considers a wireless sensor network (WSN) of a number of low-cost low-power nodes distributed across a large area for data sensing, simple information processing, and communication over short distances. In a WSN, nodes locating farther from the gateway can be assisted by nearer nodes for data delivery. Such multi-hop communications can provide significant robustness against the adverse effects of wireless shadowing and fading, allowing for broader sensing coverage, enhanced throughput and reliability compared to direct transmissions. There has been rapid adoption of WSN for demanding industrial environments thanks to recent developments in wireless communication, power efficiency, extreme miniaturization and embedded computing technologies [4], [7]– [10]. In such a WSN, the sensory data at the nodes is commonly required to be delivered to the gateway(s) or data centers by certain deadlines to provide useful and/or meaningful information. For example, in smart healthcare and smart grid applications, monitoring information has to be updated at the data center in a timely fashion to promptly evaluate or assess the health conditions of patients, or potential malfunctions and security threats to the electricity grids. Supporting delay quality-of-service (QoS) guarantees is crucial to ensure satisfactory operation of IIoT applications [11], [12].

While design and optimization of the general multi-hop WSN with delay QoS guarantees are challenging, this work considers a dual-hop network, where each node communicates the sensory information to a common gateway either directly or via another node using the dual-hop relaying mode. Nodes communicating directly with the gateway are referred to as single-hop nodes (or relays), and those indirectly dual-hop nodes (or sources). Such multi-source multi-relay networks have been extensively studied under different settings and assumptions (see, e.g., [13]- [18] and references therein). Our current work distinguishes itself from the existing works in the following important aspects. First, the relays transmit their own data in addition to the data received from other sources, which is applicable for scenarios where all nodes are sensing the environment to generate their own data. Hence, each relay has to share the capacity of the link between itself and the gateway to transmit both its own data and the data of the assisted sources. Such capacity sharing needs to take into account the QoS such as delay and rates requirements of the relay and sources, creating additional challenges in resource provisionings. Second, more importantly, this work imposes delay constraint on each source and relay sensory data in terms of a maximum acceptable *delay-outage* probability (i.e., the end-to-end delay of each source and relay data flow is allowed to exceed a delay bound within an acceptable outage probability) [19]. For many IIoT applications such as manufacturing monitoring, intelligent transport systems (ITS), smart grids, and health care, continually sensed data from sensors needs to be delivered to the gateway within a certain delay bound to be useful. However, guaranteeing a *deterministic*

[1]While the main sources of energy consumption in sensor nodes are data collection and calculation, and data transmission, the latter is much greater than the former two. Hence, this current work focuses on energy consumption in data transmission in IoT communications networks.

delay bound over fading channels requires prohibitively large power due to the possible deep fades. Consequently, this work considers delay-outage constraints, which are applicable because: 1) the mentioned applications can indeed tolerate a (small) delay bound violation probability; and 2) the need for high power consumption can be relaxed. Depending on a particular application, the delay bound and maximum allowable outage probability can be determined accordingly for satisfactory functioning. It should be emphasized that there are works on stochastic optimization of wireless networks to guarantee a maximum *average* delay bound [20]. However, guaranteeing an *average* delay bound does not guarantee a maximum *deterministic* delay bound, the latter of which is required in most IIoT applications.

This work develops energy-efficient resource allocation and transmission schemes for a dual-hop IoT network under delay-outage constraints. In the proposed model, each node maintains a transmission buffer to store its own sensory data generated at a constant rate. In addition, each relay employs a separate buffer to store the data received from each of the sources that it relays. In each transmission frame, we assume that at most one link (or node) is active to avoid strong interference. Also, when a relay–gateway link is active, the relay has to allocate different transmission rates for its own data and that of its assisted sources such that the total rates do not exceed the link capacity. For the proposed network model, the joint adaptive link scheduling, rate, and power allocation solution in each frame is determined depending on the instantaneous channel conditions, aiming to minimize the total power consumption to support minimum sensory data rate requirements and delay-outage constraints of the sources and relays.

To solve the optimization problems, the steady-state distributions of the source and relay queue (buffer) lengths need to be known to handle the delay-outage constraints. However, deriving the distributions is highly intractable. Moreover, if the large delay regime is considered (i.e., the delay bound is orders of magnitude larger than the transmission frame duration, which is typically true for practical IoT applications), we can employ the asymptotic delay analysis to compute the achieved delay exponents (or the exponential decay rates) of queue length tail distributions as such to satisfy the delay-outage constraints [21], [22]. Then, we derive an explicit relationship between the delay exponents and the resource allocation variables. The solutions to the resulting constrained optimization problems are derived using Lagrangian approach and convex optimization. Such solutions take into account the delay constraints, rate requirements, as well as the channel fading statistics through the use of the Lagrange multipliers. Moreover, we develop online transmission algorithms when the fading statistics are unknown, a typical scenario in real-life IoT networks. Numerical results are performed to demonstrate the impacts of the data rate requirements and delay constraints on the power consumption. The performance gains due to the adaptive power allocation over the fixed power allocation are also illustrated. Our results can help to design optimal routing and the corresponding resource allocation for dual-hop IoT networks to support the given rate requirements and delay-outage constraints of the nodes.

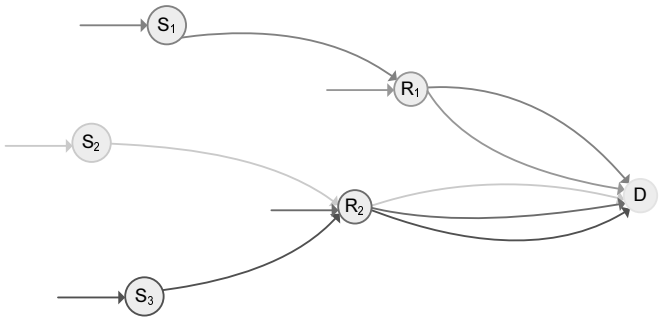


Fig. 1. A dual-hop WSN with $K = 3$ sources and $N = 2$ relays: Relay R_1 assists source S_1 and relay R_2 assists sources S_2, and S_3. Both sources and relays have own data to transmit to the destination.

The rest of the manuscript is organized as follows. Section II describes the system model and formulates the resource allocation problem with the solution approach being presented in Section III. Section IV considers the case of adaptive source and relay power allocation. Illustrative results are presented in Section V, followed by conclusions in Section VI.

II. System Model and Problem Formulation

A. Dual-hop Network Model

Consider a general dual-hop WSN consisting of $K + N$ nodes, which continually sense the environment and communicate the sensory data to a common IoT gateway (or destination) over a single channel of bandwidth B (Hz). Of these $K + N$ nodes, there are $K \geq 1$ sources $S_k, k \in \mathcal{S} = \{1, \ldots, K\}$ and $N \geq 1$ relays $R_n, n \in \mathcal{R} = \{1, \ldots, N\}$. Each source is assumed to be assisted by only one relay while one relay can assist none, one, or several sources. See Figure 1 for an example of such network. Denote by $\mathcal{S}(R_n) \subseteq \mathcal{S}$ the set of indices of the sources assisted by the relay R_n, and $n(S_k) \in \mathcal{R}$ the index of the relay assisting the source S_k. Each S_k uses a transmission buffer Q_{S_k} to store its sensory data generated at a constant rate μ_{S_k} (b/s/Hz). Each relay R_n maintains $|\mathcal{S}(R_n)| + 1$ buffers: one buffer Q_{R_n} to store its sensory data generated at a constant rate μ_{R_n} (b/s/Hz), and $|\mathcal{S}(R_n)|$ buffers $Q_{R_n}^{S_k}, k \in \mathcal{S}(R_n)$ to store the data received from the sources it assists. $Q_{S_k}[t]$, $Q_{R_n}[t]$, $Q_{R_n}^{S_k}[t]$ denote the queue lengths. and The delay of the data of source S_k is the sum queuing delays incurred at the two buffers Q_{S_k} and $Q_{R_{n(S_k)}}^{S_k}$. On the other hand, the delay of the data of relay R_n is the delay incurred at the buffer Q_{R_n} only. Table I summarizes the main notations used in the paper.

B. Wireless Transmission and Resource Allocation Model

1) Block-fading wireless channels: We assume slow block-fading channels in which channel gains remain unchanged during the transmission frame T (seconds) but can vary independently from one to another frame. Denote $h_{k,n(S_k)}[t], k \in \mathcal{S}$, and $h_{n,D}[t], n \in \mathcal{R}$ as the channel power gains in frame $t = 1, 2, \ldots$ of the S_k–$R_{n(S_k)}$ and the R_n–D links, respectively, which are assumed to be statistically independent with

TABLE I
TABLE OF MAIN NOTATIONS

Notations	Meanings
$S_k, k \in \mathcal{S} = \{1, \ldots, K\}$	K Sources
$R_n, n \in \mathcal{R} = \{1, \ldots, N\}$	N Relays
$\mathcal{S}(R_n) \subseteq \mathcal{S}$	Source indexes assisted by R_n
$n(S_k) \in \mathcal{R}$	Relay index assisting S_k
μ_{S_k}, μ_{R_n}	Constant rates of S_k and R_n
$P_{S_k}[t], P_{R_n}[t]$	Transmit powers of S_k and R_n
$Q_{S_k}[t], Q_{R_n}[t], Q_{R_n}^{S_k}[t]$	Queue lengths (in frame t)
$h_{k,n(S_k)}[t], h_{n,D}[t]$	Channel power gains
$r_{k,n(S_k)}[t], r_{n,D}[t]$	Transmission rates
$\phi_{k,n(S_k)}[t], \phi_{n,D}[t] \in \{0,1\}$	Link scheduling variables
$r_{n,D}^{S_k}[t], r_{n,D}^{R_n}[t]$	Allocated rates at R_n
$\zeta_{S_k}, \zeta_{R_n} \in (0,1]$	Delay-outage probabilities
$\theta_{S_k}, \theta_{R_n}$	Delay exponents
$\Omega_{S_k}(\theta), \Omega_{R_n}(\theta), \Omega_{R_{n(S_k)}}^{S_k}(\theta)$	Log moment generating functions

means $\mathbb{E}[h_{k,n(S_k)}]$ and $\mathbb{E}[h_{n,D}]$, where $\mathbb{E}[.]$ denotes statistical expectation operator.

Let $P_{S_k}, k \in \mathcal{S}$ and $P_{R_n}, n \in \mathcal{R}$ denote the power allocation of the sources and the relays, respectively. Using Shannon's formula, the *instantaneous* transmission rates (b/s/Hz) of the S_k–$R_{n(S_k)}$ and the R_n–D links in frame t are, respectively:

$$\begin{aligned} r_{k,n(S_k)}[t] &= \log_2\big(1 + P_{S_k} h_{k,n(S_k)}[t]\big), k \in \mathcal{S} \\ r_{n,D}[t] &= \log_2\big(1 + P_{R_n} h_{n,D}[t]\big), n \in \mathcal{R}. \end{aligned} \tag{1}$$

2) Resource allocation model: We now describe the link scheduling and rate allocation problem.

(a) Link scheduling. In frame t, let $\phi_{k,n(S_k)}[t], k \in \mathcal{S}$ and $\phi_{n,D}[t], n \in \mathcal{R}$ denote the binary scheduling variables for the S_k–$R_{n(S_k)}$ and the R_n–D links, respectively. The scheduling variable is set to 1 if the corresponding link is active, otherwise, 0. We assume that at most one link is allowed to be active in each frame t to avoid strong interference. Hence, we have:

$$\begin{aligned} &\phi_{k,n(S_k)}[t], \phi_{n,D}[t] \in \{0,1\}, \forall t, k \in \mathcal{S}, n \in \mathcal{R} \\ &\sum_{k \in \mathcal{S}} \phi_{k,n(S_k)}[t] + \sum_{n \in \mathcal{R}} \phi_{n,D}[t] \leq 1, \forall t. \end{aligned} \tag{2}$$

Note that it is possible that no link is active in a frame in order to save power due to unfavorable channel conditions. The average total network power is computed as:

$$\mathcal{P}_{\text{avg}} = \mathbb{E}\Big[\sum_{k \in \mathcal{S}} \phi_{k,n(S_k)}[t] P_{S_k} + \sum_{n \in \mathcal{R}} \phi_{n,D}[t] P_{R_n}\Big]. \tag{3}$$

(b) Rate allocation: In frame t, if the S_k–$R_{n(S_k)}$ link is active, source S_k transmits its currently data in queue Q_{S_k} to relay $R_{n(S_k)}$ at a rate $r_{k,n(S_k)}[t]$ in (1). If the R_n–D link is active, relay R_n has to solve the rate allocation problem by determining the rates $r_{n,D}^{S_k}[t] \geq 0, k \in \mathcal{S}(R_n)$ and $r_{n,D}^{R_n}[t] \geq 0$ to transmit respectively the data of its assisted sources in queue $Q_{R_n}^{S_k}[t]$ and that of its own data in queue $Q_{R_n}[t]$ to the destination so that the total allocated rates do not exceed the link capacity $r_{n,D}[t]$. The rate allocation constraint at the relay R_n is then expressed as:

$$\begin{aligned} &r_{n,D}^{S_k}[t], r_{n,D}^{R_n}[t] \geq 0, \forall t, k \in \mathcal{S}(R_n), \\ &\sum_{k \in \mathcal{S}(R_n)} r_{n,D}^{S_k}[t] + r_{n,D}^{R_n}[t] \leq \phi_{n,D}[t] r_{n,D}[t], \forall t. \end{aligned} \tag{4}$$

3) Queue dynamics: The dynamics of the queue lengths (of all queues) at all nodes (including sources and relays) from frame t to frame $t+1$ are given as:

$$\begin{aligned} Q_{S_k}[t+1] &= Q_{S_k}[t] - \min\big\{Q_{S_k}[t], \phi_{k,n(S_k)}[t] TB \\ &\quad r_{k,n(S_k)}[t]\big\} + \mu_{S_k} TB, k \in \mathcal{S} \\ Q_{R_n}[t+1] &= Q_{R_n}[t] - \min\big\{Q_{R_n}[t], \phi_{n,D}[t] TB \\ &\quad r_{n,D}^{R_n}[t]\big\} + \mu_{R_n} TB, n \in \mathcal{R} \\ Q_{R_{n(S_k)}}^{S_k}[t+1] &= Q_{R_{n(S_k)}}^{S_k}[t] - \min\big\{Q_{R_{n(S_k)}}^{S_k}[t], \phi_{n(S_k),D}[t] \\ &\quad TB r_{n(S_k),D}^{S_k}[t]\big\} + \min\big\{Q_{S_k}[t], \phi_{k,n(S_k)}[t] \\ &\quad TB r_{k,n(S_k)}[t]\big\}, k \in \mathcal{S}. \end{aligned} \tag{5}$$

In the right hand side of each expression, the second and third terms are the data amounts removed from, and arriving to the buffer in frame t, respectively. The data amount removed from a queue is limited by the current data amount buffered [21]. It can be seen that the dynamics and the steady states of the queue lengths, and hence the delay of buffered data are governed by the underlying resource allocation solutions/schemes.

C. Delay-outage Constraints

We assume stable queues. As $t \to \infty$, the random queue length processes in (5) converge to steady-state queue length random variables also denoted as Q_{S_k}, $Q_{R_{n(S_k)}}^{S_k}$, and Q_{R_n} for simplicity. The delay-outage constraints for the sources and relays are expressed as:

$$\begin{aligned} \Pr\big(Q_{S_k} + Q_{R_{n(S_k)}}^{S_k} > Q^{\max}\big) \leq \zeta_{S_k}, k \in \mathcal{S} \\ \Pr\big(Q_{R_n} > Q^{\max}\big) \leq \zeta_{R_n}, n \in \mathcal{R} \end{aligned} \tag{6}$$

where $\Pr(x > y)$ denotes the probability of the event $x > y$; $Q^{\max}$ denotes the queue length bound, which is assumed to be the same for all sources and relays without loss of generality; ζ_{S_k}, $\zeta_{R_n} \in (0,1]$ are the maximum acceptable outage probabilities. For a given bound $Q^{\max}$, the smaller the maximum outage probability, the more stringent the delay constraint is. In addition, as the outage probability approaches 1, we allow unconstrained queue length (or delay). On the other hand, as the outage probability approaches 0, the delay outage is not allowed to happen (i.e., a bounded delay constraint). Such delay-outage constraint model can be used to model a vast number of delay-sensitive applications with diverse delay requirements.

D. Resource Allocation Problem

The resource allocation problem is formulated as:

$$\begin{aligned} \min \quad & \mathcal{P}_{\text{avg}} \\ \texttt{s.t.:} \quad & \text{Constraints } (2),(4),(5),(6) \\ & \mu_{S_k} \geq \mu_{S_k}^{\min}, k \in \mathcal{S} \\ & \mu_{R_n} \geq \mu_{R_n}^{\min}, n \in \mathcal{R} \end{aligned} \tag{7}$$

with the optimization variables $\phi_{k,n(S_k)}[t]$, $\phi_{n,D}[t]$, $r_{n,D}^{S_k}[t]$, and $r_{n,D}^{R_n}[t]$; $\mu_{S_k}^{\min}$ and $\mu_{R_n}^{\min}$ are the minimum data rate requirements for the sources and relays, respectively.

Remark 1: At optimality of (7), we must have:

i) the inequality rate constraints are met with equalities. Otherwise, more power is required to support unnecessarily higher rates.

ii) the delay-outage constraints (6) are met with equalities. Otherwise, more power is required to support unnecessarily more stringent delay constraints [21].

iii) when one or more of the outage probabilities approaches 0, (7) becomes infeasible because it requires infinitely large power to support a given positive rate with a deterministic delay bound constraint under fading channels [21].

Remark 2: To implement the resource allocation schemes under consideration, it is implicitly assumed that a centralized coordinator knows the channel conditions in each frame. It then determines the allocation solution and informs the source or the relay to be transmitting only. Such a computer can be embedded into the IoT gateway.

The following section will present the solution approach to (7) using asymptotic delay analysis.

III. Resource allocation solution

To solve (7), one usually needs to know the tail distributions of steady-state queue length random variables in (5), which is very difficult in general. To circumvent this problem, we employ the asymptotic delay analysis results to characterize the tail distributions of the queue lengths with the assumption of large queue length (or delay) regime, i.e., the bound $Q^{\max}$ is sufficiently large but finite as compared with the average frame arrival rates (i.e., $\mu_{S_k}TB$ or $\mu_{R_n}TB$).

A. Asymptotic Delay Analysis

1) Background: Consider a stable queue with an infinite buffer size, a stationary ergodic arrival process $a[t]$, and a service process $c[t], t = 1, 2, \ldots$ satisfying the Gartner-Ellis limit (i.e., for all $\theta \geq 0$, their differential asymptotic logarithmic moment generating functions (LMGFs) $\Omega_{\text{a}}(\theta)$ and $\Omega_{\text{c}}(\theta)$ exist). If there exists a unique achieved delay exponent $\bar{\theta} > 0$ satisfying:

$$\Omega_{\text{a}}(\bar{\theta}) + \Omega_{\text{c}}(-\bar{\theta}) = 0 \tag{8}$$

then for a sufficiently large x, the following result for the tail distribution of the steady-state queue length $Q_{\text{a,c}}$ holds [22]:

$$\Pr\big(Q_{\text{a,c}} > x\big) = e^{-\bar{\theta}x}. \tag{9}$$

Note that for independent and identically distributed (i.i.d.) processes, the LMGFs are computed as:

$$\Omega_{\text{a}}(\theta) = \log \mathbb{E}\big[e^{\theta a[t]}\big], \quad \Omega_{\text{c}}(\theta) = \log \mathbb{E}\big[e^{\theta c[t]}\big]. \tag{10}$$

2) Application: We now show how to apply the above results to determine the achieved delay exponents of the source and relay queues so that the constraints (6) are satisfied.

Case 1. Consider the queue Q_{R_n} for relay R_n with achieved delay exponent $\bar{\theta}_{R_n}$. To satisfy (6) for R_n, it is required that:

$$\bar{\theta}_{R_n} \geq \theta_{R_n}^{\text{tar}} \triangleq -\log(\zeta_{R_n})/Q^{\max} \tag{11}$$

as from (9), we would have:

$$\Pr\big(Q_{R_n} > Q^{\max}\big) = e^{-\bar{\theta}_{R_n}Q^{\max}} \leq e^{-\theta_{R_n}^{\text{tar}}Q^{\max}} = \zeta_{R_n}.$$

Case 2. Consider two tandem queues Q_{S_k} and $Q_{R_{n(S_k)}}^{S_k}$ for source S_k with achieved delay exponents $\bar{\theta}_{S_k}$ and $\bar{\theta}_{R_{n(S_k)}}^{S_k}$ respectively.

In [21], it has been shown that an optimal resource allocation must guarantee similar queue length distributions at both queues, implying $\bar{\theta}_{S_k} = \bar{\theta}_{R_{n(S_k)}}^{S_k}$. Moreover, to satisfy (6) for S_k, we must have:

$$\bar{\theta}_{S_k} = \bar{\theta}_{R_{n(S_k)}}^{S_k} \geq \theta_{S_k}^{\text{tar}} \triangleq -\frac{1}{Q^{\max}}\Big(1 + \mathcal{W}_{-1}\Big(-\frac{\zeta_{S_k}}{e}\Big)\Big) \tag{12}$$

where $\mathcal{W}_{-1}(.)$ is the Lambert W function's real brunch for the range $(-\infty, -1]$, and the Lambert W function is the inverse function of $Z(W) = We^W$.

In *Remark 1*, we mentioned that in order to achieve the smallest power, the constraints (6) must be met with equalities, and hence, the achieved delay exponents must be equal to the target delay exponents in (11) and (12).

B. Problem Reformulation

Previously, we have computed the achieved delay exponents of the queues of the sources and relays to satisfy (6). We now apply the relation (8) with corresponding arrival and service processes for all queues.

Consider the source S_k with a constant arrival rate μ_{S_k} into its queue Q_{S_k}. By applying the relation (8) to the (tandem) queues Q_{S_k} and $Q_{R_{n(S_k)}}^{S_k}$ with $\bar{\theta}_{S_k} = \bar{\theta}_{R_{n(S_k)}}^{S_k} = \theta_{S_k}^{\text{tar}}$, we have:

$$\begin{aligned} \mu_{S_k}TB\bar{\theta}_{S_k} + \Omega_{S_k}(-\bar{\theta}_{S_k}) &= 0 \\ \Omega_{R_{n(S_k)}}^{S_k,\text{arv}}(\bar{\theta}_{R_{n(S_k)}}^{S_k}) + \Omega_{R_{n(S_k)}}^{S_k}(-\bar{\theta}_{R_{n(S_k)}}^{S_k}) &= 0. \end{aligned} \tag{13}$$

$\Omega_{S_k}(\theta)$ and $\Omega_{R_{n(S_k)}}^{S_k}(\theta)$ are the LMGFs of the service processes of the queues Q_{S_k} and $Q_{R_{n(S_k)}}^{S_k}$, which can be computed using (10) as follows:

$$\begin{aligned} \Omega_{S_k}(\theta) &= \log \mathbb{E}\big[e^{\theta\phi_{k,n(S_k)}[t]TBr_{k,n(S_k)}[t]}\big] \\ \Omega_{R_{n(S_k)}}^{S_k}(\theta) &= \log \mathbb{E}\big[e^{\theta\phi_{n(S_k),D}[t]TBr_{n(S_k),D}^{S_k}[t]}\big] \end{aligned} \tag{14}$$

using the resource allocation variables described in Section II. Note that the LMGF of the constant arrival process with rate μ_{S_k} to Q_{S_k} is $\mu_{S_k}TB\theta$ (see (10)). Also, $\Omega_{R_{n(S_k)}}^{S_k,\text{arv}}(\theta)$ is the LMGF of the arrival process at the queue $Q_{R_{n(S_k)}}^{S_k}$, which is the same as the LMGF of the departure process of Q_{S_k} and is given by [21]:

$$\Omega_{R_{n(S_k)}}^{S_k,\text{arv}}(\theta) = \begin{cases} \mu_{S_k}TB\theta, & 0 \leq \theta \leq \bar{\theta}_{S_k}, \\ \mu_{S_k}TB\bar{\theta}_{S_k} + \Omega_{S_k}(\theta - \bar{\theta}_{S_k}), & \theta > \bar{\theta}_{S_k}. \end{cases} \tag{15}$$

We then have:

$$\Omega^{S_k,\text{arv}}_{R_{n(S_k)}}(\bar{\theta}^{S_k}_{R_{n(S_k)}}) = \mu_{S_k} TB \bar{\theta}^{S_k}_{R_{n(S_k)}}.$$

Now consider the relay R_n with a constant arrival rate μ_{R_n} into its queue Q_{R_n} with $\bar{\theta}_{R_n} = \theta^{\text{tar}}_{R_n}$. Again, by applying the relation (8) to Q_{R_n}, we have:

$$\mu_{R_n} TB \bar{\theta}_{R_n} + \Omega_{R_n}(-\bar{\theta}_{R_n}) = 0 \tag{16}$$

$\Omega_{R_n}(\theta)$ is the LMGF of the service process of Q_{R_n}, which is computed using (10) as:

$$\Omega_{R_n}(\theta) = \log \mathbb{E}\big[e^{\theta \phi_{n,D}[t] TB r^{R_n}_{n,D}[t]}\big]. \tag{17}$$

The LMGF of the constant arrival process to Q_{R_n} is $\mu_{R_n} TB\theta$ (see (10)).

We have argued that $\mu_{S_k} = \mu^{\min}_{S_k}, k \in \mathcal{S}$, and $\mu_{R_n} = \mu^{\min}_{R_n}, n \in \mathcal{R}$. Using (13), (16), the problem (7) can now be reformulated as:

$$\begin{aligned} \min \quad & \mathcal{P}_{\text{avg}} \\ \texttt{s.t.:} \quad & \text{Constraints } (2), (4), \\ & \mu^{\min}_{S_k} TB \theta^{\text{tar}}_{S_k} + \Omega_{S_k}(-\theta^{\text{tar}}_{S_k}) = 0, k \in \mathcal{S}, \\ & \mu^{\min}_{S_k} TB \theta^{\text{tar}}_{S_k} + \Omega^{S_k}_{R_{n(S_k)}}(-\theta^{\text{tar}}_{S_k}) = 0, k \in \mathcal{S}, \\ & \mu^{\min}_{R_n} TB \theta^{\text{tar}}_{R_n} + \Omega_{R_n}(-\theta^{\text{tar}}_{R_n}) = 0, n \in \mathcal{R}. \end{aligned} \tag{18}$$

Using (14), (17), and the monotonicity of log function, after some manipulations, we can re-write (18) in terms of the resource allocation variables as:

$$\begin{aligned} \min \quad & \mathbb{E}\Big[\sum_{k\in\mathcal{S}} \phi_{k,n(S_k)}[t] P_{S_k} + \sum_{n\in\mathcal{R}} \phi_{n,D}[t] P_{R_n}\Big] \\ \texttt{s.t.:} \quad & \text{Constraints } (2), (4), \\ & \mathbb{E}\big[e^{-\theta_{S_k} \phi_{n(S_k),D}[t] r^{S_k}_{n(S_k),D}[t]}\big] = \\ & \qquad \mathbb{E}\big[e^{-\theta_{S_k} \phi_{k,n(S_k)}[t] r_{k,n(S_k)}[t]}\big], k \in \mathcal{S}, \\ & \mathbb{E}\big[e^{-\theta_{S_k} \phi_{k,n(S_k)}[t] r_{k,n(S_k)}[t]}\big] = e^{-\mu^{\min}_{S_k} \theta_{S_k}}, k \in \mathcal{S}, \\ & \mathbb{E}\big[e^{-\theta_{R_n} \phi_{n,D}[t] r^{R_n}_{n,D}[t]}\big] = e^{-\mu^{\min}_{R_n} \theta_{R_n}}, n \in \mathcal{R} \end{aligned} \tag{19}$$

where for notational convenience, we have defined the following normalized delay exponents:

$$\theta_{S_k} \triangleq \theta^{\text{tar}}_{S_k} TB, k \in \mathcal{S}, \quad \theta_{R_n} \triangleq \theta^{\text{tar}}_{R_n} TB, n \in \mathcal{R}.$$

The equality constraints in (19) can be equivalently replaced by inequality $\leq$ constraints without loosing optimality. The remaining task is to solve (19), which is described below.

C. Resource Allocation Solution via Lagrangian Approach

We employ the Lagrangian approach to solve (19). We form the partial Lagrangian of (19) as:

$$\mathcal{L} = \mathbb{E}\big[\mathcal{L}[t]\big] \tag{20}$$

with

$$\begin{aligned} \mathcal{L}[t] \;=\; & \sum_{k\in\mathcal{S}} \phi_{k,n(S_k)}[t] P_{S_k} + \sum_{n\in\mathcal{R}} \phi_{n,D}[t] P_{R_n} \\ & + \sum_{k\in\mathcal{S}} \Big((\omega_k - \lambda_k) e^{-\theta_{S_k} \phi_{k,n(S_k)}[t] r_{k,n(S_k)}[t]} \\ & + \lambda_k e^{-\theta_{S_k} \phi_{n(S_k),D}[t] r^{S_k}_{n(S_k),D}[t]} \Big) + \sum_{n\in\mathcal{R}} \xi_n e^{-\theta_{R_n} \phi_{n,D}[t] r^{R_n}_{n,D}[t]} \end{aligned} \tag{21}$$

where $\lambda_k, \omega_k \geq 0, k \in \mathcal{S}$ and $\xi_n \geq 0, n \in \mathcal{R}$ are the Lagrange multipliers corresponding to the inequality constraints in (19).

Now, if we minimize the Lagrangian $\mathcal{L}$ under the constraints (2), (4), and the multipliers are determined to satisfy the inequality constraints in (19) with equalities, we would obtain the solution of (19). Moreover, in order to minimize $\mathcal{L}$, the optimal link scheduling and rate allocation solution has to minimize $\mathcal{L}[t]$ in each frame $t = 1, 2, \ldots$ as

$$\min \;\; \mathcal{L}[t] \quad \texttt{s.t.:} \;\; \text{Constraints } (2), (4). \tag{22}$$

The problem (22) is a mixed-integer optimization problem. The solution can be obtained by considering the following three cases:

Case 1. Consider $\phi_{k,n(S_k)}[t] = 1, k \in \mathcal{S}$. The objective value of (22) (or the scheduling metric) is:

$$\begin{aligned} \mathcal{L}[t]\Big|_{\phi_{k,n(S_k)}[t]=1} \;=\; & P_{S_k} + (\omega_k - \lambda_k) e^{-\theta_{S_k} r_{k,n(S_k)}[t]} \\ & + \lambda_k + \sum_{k' \neq k} \omega_{k'} + \sum_{n\in\mathcal{R}} \xi_n. \end{aligned} \tag{23}$$

In this case, there is no rate allocation problem to address.

Case 2. Consider $\phi_{n,D}[t] = 1, n \in \mathcal{R}$. We have the following rate allocation problem at relay R_n:

$$\begin{aligned} \min \quad & \sum_{k\in\mathcal{S}(R_n)} \lambda_k e^{-\theta_{S_k} r^{S_k}_{n,D}[t]} + \xi_n e^{-\theta_{R_n} r^{R_n}_{n,D}[t]} \\ \texttt{s.t.:} \quad & \sum_{k\in\mathcal{S}(R_n)} r^{S_k}_{n,D}[t] + r^{R_n}_{n,D}[t] \leq r_{n,D}[t] \end{aligned} \tag{24}$$

where recall that $\mathcal{S}(R_n)$ is the set of indices of the sources assisted by R_n.

It can be verified that (24) is a convex optimization problem due to its convex objective function and constraint. We can further employ the Lagrangian approach to determine the optimal rate allocation solution by solving the following optimization problem:

$$\begin{aligned} \min \quad & \sum_{k\in\mathcal{S}(R_n)} \lambda_k e^{-\theta_{S_k} r^{S_k}_{n,D}[t]} + \xi_n e^{-\theta_{R_n} r^{R_n}_{n,D}[t]} \\ & + \chi_{n,D}[t] \Big(\sum_{k\in\mathcal{S}(R_n)} r^{S_k}_{n,D}[t] + r^{R_n}_{n,D}[t] \Big) \end{aligned} \tag{25}$$

where $\chi_{n,D}[t] > 0$ is some Lagrange multiplier such that the rate inequality constraint in (24) is met with equality. By differentiating the objective function, setting it to zero, and accounting for the non-negativeness of the data rates, the rate allocation solution is derived as:

$$\begin{aligned} r^{S_k}_{n,D}[t] &= \left[-\frac{1}{\theta_{S_k}} \log\Big(\frac{\chi_{n,D}[t]}{\lambda_k \theta_{S_k}} \Big) \right]^+, k \in \mathcal{S}(R_n), \\ r^{R_n}_{n,D} &= \left[-\frac{1}{\theta_{R_n}} \log\Big(\frac{\chi_{n,D}[t]}{\xi_n \theta_{R_n}} \Big) \right]^+ \end{aligned} \tag{26}$$

where $[x]^+$ denotes $\max\{x, 0\}$. Then, we can solve for $\chi_{n,D}[t]$ to satisfy (24) with equality using numerical methods. We omit

the details for brevity. The objective function of (22) in the case $\phi_{n,D}[t]=1$ is then

$$\mathcal{L}[t]\Big|_{\phi_{n,D}[t]=1} = P_{R_n} + \sum_{k\in\mathcal{S}}(\omega_k-\lambda_k) + \sum_{k\notin\mathcal{S}(R_n)}\lambda_k + \sum_{k\in\mathcal{S}(R_n)}\lambda_k e^{-\theta_{S_k} r_{n,D}^{S_k}[t]} + \xi_n e^{-\theta_{R_n} r_{n,D}^{R_n}[t]} + \sum_{n'\neq n}\xi_{n'}. \quad (27)$$

Case 3. Consider $\phi_{k,n(S_k)}[t]=\phi_{n,D}[t]=0, k\in\mathcal{S}, n\in\mathcal{R}$. In this case, no link is active. The objective function of (22) becomes:

$$\mathcal{L}[t]\Big|_{\phi_{k,n(S_k)}[t]=\phi_{n,D}[t]=0,k\in\mathcal{S},n\in\mathcal{R}} = \sum_{k\in\mathcal{S}}\omega_k + \sum_{n\in\mathcal{R}}\xi_n. \quad (28)$$

In summary, the optimal link scheduling and rate allocation in each frame t is determined to provide the smallest scheduling metric (23), (27), or (28). The scheduling complexity is linear in the total number of sources and relays $K+N$.

D. Online Iterative Algorithm under Unknown Fading Statistics

The resource allocation solution depends on the multipliers λ_k, $\omega_k, k\in\mathcal{S}$ and $\xi_n, n\in\mathcal{R}$, which satisfy the equality constraints in (19). To compute these multipliers, we need to know the fading statistics, which is often unavailable in reality. Even when the statistics is known, it is not easy to compute the expectations in (19). To overcome these difficulties, we can utilize the following stochastic iterations. Specifically, we first initialize the multipliers with $\lambda_k[1]>0, \omega_k[1]>0, k\in\mathcal{S}$, and $\xi_n[1]>0, n\in\mathcal{R}$. Then, in transmission frame $t=1,2,\ldots$, we carry out the following updates:

$$\lambda_k[t+1] = \Big[\lambda_k[t] + \epsilon[t]\Big(e^{-\theta_{S_k}\phi_{n(S_k),D}[t] r_{n(S_k),D}^{S_k}[t]} - e^{-\theta_{S_k}\phi_{k,n(S_k)}[t] r_{k,n(S_k)}[t]}\Big)\Big]_{\varepsilon}^{L}, k\in\mathcal{S}$$

$$\omega_k[t+1] = \Big[\omega_k[t] + \epsilon[t]\Big(e^{-\theta_{S_k}\phi_{k,n(S_k)}[t] r_{k,n(S_k)}[t]} - e^{-\mu_{S_k}^{\min}\theta_{S_k}}\Big)\Big]_{\varepsilon}^{L}, k\in\mathcal{S}$$

$$\xi_n[t+1] = \Big[\xi_n[t] + \epsilon[t]\Big(e^{-\theta_{R_n}\phi_{n,D}[t] r_{n,D}^{R_n}[t]} - e^{-\mu_{R_n}^{\min}\theta_{R_n}}\Big)\Big]_{\varepsilon}^{L}, n\in\mathcal{R}$$

for some small coefficient $\varepsilon>0$. Here, $[x]_a^b$ denotes the projection of x on the interval $[a,b]$ for $b\geq a\geq 0$ and L is sufficiently large to ensure boundedness of the multiplier updates. The decreasing positive sequence $\epsilon[t]$ which dictates the convergence speed, must satisfy:

$$\sum_{t=1}^{\infty}\epsilon[t]=\infty; \quad \sum_{t=1}^{\infty}(\epsilon[t])^2<\infty.$$

The link scheduling and rate allocation solution in frame t is computed using the current estimates of the Lagrange multipliers. We can see that these updates do not require the fading statistical knowledge. Moreover, the algorithm does not assume any specification on the fading statistics, and it converges for any independent link fading distributions. The convergence and optimality of the stochastic iteration based online algorithm can be established using the results in stochastic approximation theory. The details of convergence proof, though highly technical and lengthy, are routine and hence, omitted due to lack of space. More interested readers are referred to [23], [24] for the details of the convergence proof for similar algorithms, albeit under different settings.

IV. Resource allocation with adaptive power allocation

In previous sections, we have assumed a fixed source and relay power allocation $P_{S_k}, k\in\mathcal{S}$ and $P_{R_n}, n\in\mathcal{R}$ for all frames. It is known that adaptive power adaption can be employed to exploit the temporal fading diversity for power savings. Denote $P_{S_k}[t], k\in\mathcal{S}$ and $P_{R_n}[t], n\in\mathcal{R}$ as the power allocation of the source S_k and relay R_n in frame t, respectively. The average total network power is given by:

$$\mathcal{P}_{\text{avg}}^{\text{APA}} = \mathbb{E}\Big[\sum_{k\in\mathcal{S}}\phi_{k,n(S_k)}[t]P_{S_k}[t] + \sum_{n\in\mathcal{R}}\phi_{n,D}[t]P_{R_n}[t]\Big]. \quad (29)$$

$P_{S_k}[t]$ (or $P_{R_n}[t]$) can only be positive when $\phi_{k,n(S_k)}[t]=1$ (or $\phi_{n,D}[t]=1$).

A. Problem Reformulation

After some manipulations, similar to (19), the resource allocation problem with adaptive power allocation can be expressed as:

$$\begin{aligned}
\min\quad & \mathbb{E}\Big[\sum_{k\in\mathcal{S}}\phi_{k,n(S_k)}[t]P_{S_k}[t] + \sum_{n\in\mathcal{R}}\phi_{n,D}[t]P_{R_n}[t]\Big] \\
\texttt{s.t.:}\ & \text{Constraints } (2),(4) \\
& \mathbb{E}\big[e^{-\theta_{S_k}\phi_{n(S_k),D}[t] r_{n(S_k),D}^{S_k}[t]}\big] \leq \\
& \mathbb{E}\big[\big(1+h_{k,n(S_k)}[t]P_{S_k}[t]\big)^{-\theta_{S_k}\phi_{k,n(S_k)}[t]/\log(2)}\big], k\in\mathcal{S} \\
& \mathbb{E}\big[\big(1+h_{k,n(S_k)}[t]P_{S_k}[t]\big)^{-\theta_{S_k}\phi_{k,n(S_k)}[t]/\log(2)}\big] \\
& \qquad \leq e^{-\mu_{S_k}^{\min}\theta_{S_k}}, k\in\mathcal{S} \\
& \mathbb{E}\big[e^{-\theta_{R_n}\phi_{n,D}[t] r_{n,D}^{R_n}[t]}\big] \leq e^{-\mu_{R_n}^{\min}\theta_{R_n}}, n\in\mathcal{R}.
\end{aligned} \quad (30)$$

The optimization variables are $\phi_{k,n(S_k)}[t]$, $\phi_{n,D}[t]$, $r_{n(S_k),D}^{S_k}[t]$, $r_{n,D}^{R_n}[t]$, $P_{S_k}[t]$, and $P_{R_n}[t]$.

B. Optimal Solution

Similar to (20), (21), the Lagrangian of (30) can be formed:

$$\mathcal{L} = \mathbb{E}\big[\mathcal{L}[t]\big] \quad (31)$$

where

$$\begin{aligned}
\mathcal{L}[t] = & \sum_{k\in\mathcal{S}}\phi_{k,n(S_k)}[t]P_{S_k}[t] + \sum_{k\in\mathcal{S}}\Big((\omega_k-\lambda_k) \\
& \big(1+h_{k,n(S_k)}[t]P_{S_k}[t]\big)^{-\theta_{S_k}\phi_{k,n(S_k)}[t]/\log(2)} \\
& +\lambda_k e^{-\theta_{S_k}\phi_{n(S_k),D}[t] r_{n(S_k),D}^{S_k}[t]}\Big) \\
& + \sum_{n\in\mathcal{R}}\phi_{n,D}[t]P_{R_n}[t] + \sum_{n\in\mathcal{R}}\xi_n e^{-\theta_{R_n}\phi_{n,D}[t] r_{n,D}^{R_n}[t]}.
\end{aligned}$$

Here, we have re-used the notations for the Lagrange multipliers in Section III for simplicity since it would not cause any ambiguity.

To minimize the Lagrangian $\mathcal{L}$ in (31), we have to find the link scheduling, rate, and power allocation solution in each frame t to minimize $\mathcal{L}[t]$ as

$$\min \quad \mathcal{L}[t] \quad \texttt{s.t.:} \text{ Constraints } (2), (4). \tag{32}$$

To solve (32), we consider the following cases.

Case 1. Consider $\phi_{k,n(S_k)}[t] = 1, k \in \mathcal{S}$. We solve the following power allocation problem to find $P_{S_k}[t]$:

$$\min_{P_{S_k}[t]} \; P_{S_k}[t] + (\omega_k - \lambda_k)\big(1 + h_{k,n(S_k)}[t] P_{S_k}[t]\big)^{-\theta_{S_k}/\log(2)}. \tag{33}$$

It can be verified that (33) is a convex optimization problem. Thus, by differentiating the objective function, setting it equal to 0 and accounting for the non-negativeness of the power variables, the optimal power allocation solution is derived as:

$$P_{S_k}[t] = \frac{1}{h_{k,n(S_k)}[t]} \times \left[\left((\omega_k - \lambda_k)\frac{\theta_{S_k}}{\log(2)} h_{k,n(S_k)}[t]\right)^{\left(\theta_{S_k}/\log(2)+1\right)^{-1}} - 1\right]^+. \tag{34}$$

Then, the objective function value $\mathcal{L}[t]\Big|_{\phi_{k,n(S_k)}[t]=1}$ in (32) can be computed.

Case 2. Consider $\phi_{n,D}[t] = 1, n \in \mathcal{R}$. We have the following joint rate and power allocation problem at R_n:

$$\begin{aligned} \min_{r_{n,D}^{S_k}[t], r_{n,D}^{R_n}[t], P_{R_n}[t]} \quad & P_{R_n}[t] + \sum_{k \in \mathcal{S}(R_n)} \lambda_k e^{-\theta_{S_k} r_{n,D}^{S_k}[t]} \\ & + \xi_n e^{-\theta_{R_n} r_{n,D}^{R_n}[t]} \\ \texttt{s.t.:} \quad & \sum_{k \in \mathcal{S}(R_n)} r_{n,D}^{S_k}[t] + r_{n,D}^{R_n}[t] \\ & \le \log_2\Big(1 + h_{n,D}[t] P_{R_n}[t]\Big). \end{aligned} \tag{35}$$

It can be shown that (35) is a convex optimization problem, which can be solved using Lagrangian approach. The rate allocation solution is computed as follows:

$$\begin{aligned} r_{n,D}^{S_k}[t] &= \left[-\frac{1}{\theta_{S_k}} \log\Big(\frac{\chi_{n,D}[t]}{\lambda_k \theta_{S_k}}\Big)\right]^+, k \in \mathcal{S}(R_n), \\ r_{n,D}^{R_n}[t] &= \left[-\frac{1}{\theta_{R_n}} \log\Big(\frac{\chi_{n,D}[t]}{\xi_n \theta_{R_n}}\Big)\right]^+ \end{aligned} \tag{36}$$

for some Lagrange multiplier $\chi_{n,D}[t] > 0$. The power allocation solution is derived as:

$$P_{R_n}[t] = \left[\frac{\chi_{n,D}[t]}{\log(2)} - \frac{1}{h_{n,D}[t]}\right]^+. \tag{37}$$

The multiplier $\chi_{n,D}[t]$ is determined such that the inequality rate constraint in (35) is met with equality. We omit the details for brevity.

We can then compute the objective function value $\mathcal{L}[t]\Big|_{\phi_{n,D}[t]=1}$ in (32).

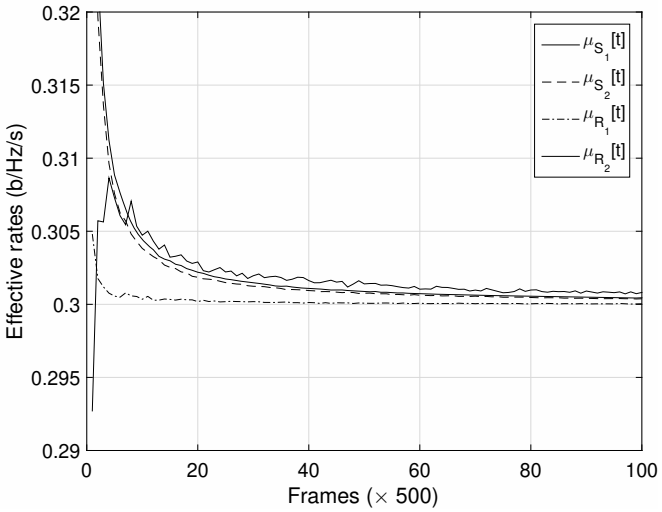


Fig. 2. Convergence of the online iterative algorithm.

Case 3. Consider $\phi_{k,n(S_k)}[t] = \phi_{n,D}[t] = 0, k \in \mathcal{S}, n \in \mathcal{R}$. The objective function value of (32) in this case can be computed as:

$$\mathcal{L}[t]\Big|_{\phi_{k,n(S_k)}[t]=\phi_{n,D}[t]=0, k\in\mathcal{S}, n\in\mathcal{R}} = \sum_{k\in\mathcal{S}} \omega_k + \sum_{n\in\mathcal{R}} \xi_n. \tag{38}$$

In summary, in frame t, the link scheduling, rate, and power allocation solution is determined as such to provide minimum objective function value of (32). Again, the multipliers λ_k, $\omega_k, k \in \mathcal{S}$, and $\xi_n, n \in \mathcal{R}$ are determined such that the associated inequality constraints in (30) are met with equalities.

V. Illustrative Results

A. Simulation Setup

Consider a narrow-band IoT (NB-IoT) dual-hop network operating over LTE bandwidth of $B = 180$ kHz with $K = 2$ sources S_1 and S_2, and $N = 2$ relays R_1 and R_2. Assume Rayleigh fading channels with a block-fading duration $T = 5$ ms. Assume the distances between nodes S_1–R_1, S_2–R_2, R_1–D, and R_2–D are equal, and are twice farther than the distances between nodes S_1–R_2, and S_2–R_1, which are assumed to be equal. To model the relative received link signal strengths, we use a path loss model with a path loss exponent of 2.5. The average link channel power gains are set as follows: $\mathbb{E}[h_{1,1}] = \mathbb{E}[h_{2,2}] = \mathbb{E}[h_{1,D}] = \mathbb{E}[h_{2,D}] = \text{SNR}_0 = 30$ dB, and hence, $\mathbb{E}[h_{1,2}] = \mathbb{E}[h_{2,1}] = \text{SNR}_0 - 10\log_{10}(2^{2.5})$ dB. There are four possible network configurations:

C1: both S_1 and S_2 are assisted by relay R_1;
C2: S_1 and S_2 are assisted by R_1 and R_2, respectively;
C3: S_1 and S_2 are assisted by R_2 and R_1, respectively;
C4: both S_1 and S_2 are assisted by R_2.

We assume homogeneous sources with similar rate requirements and delay constraints. Hence, configuration C3 will require higher power than configuration C2, and thus, C3 will not be considered in the following numerical studies since it is certainly not optimal. We denote $\mathcal{P}_{\text{avg}}^{R_1,R_1}$, $\mathcal{P}_{\text{avg}}^{R_1,R_2}$, and $\mathcal{P}_{\text{avg}}^{R_2,R_2}$ as the average powers for configurations C1, C2, and C4, respectively.

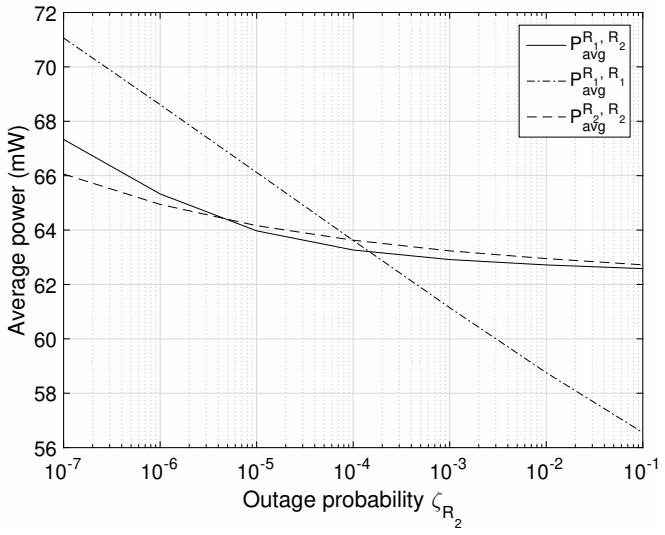


Fig. 3. Average powers versus ζ_{R_2} for different configurations.

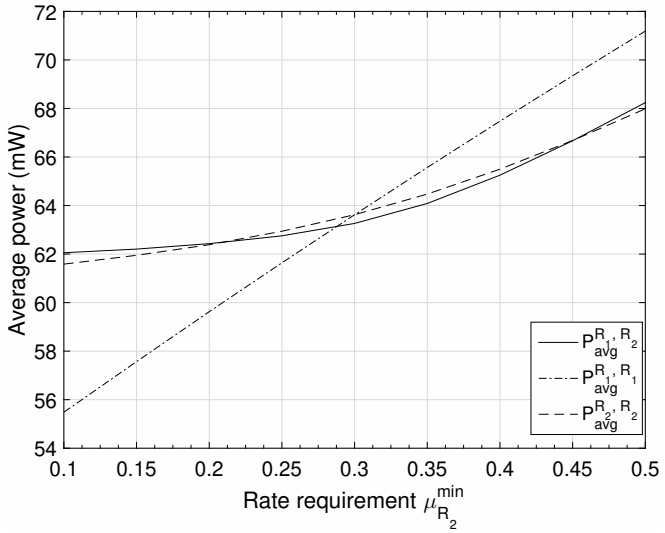


Fig. 4. Average powers versus $\mu_{R_2}^{\min}$ for different configurations.

For the delay constraints (6), we fix $Q^{\max} = 20$ kb to obtain the numerical results. The rate requirement for each node (sources and relays) varies from 0.1 b/s/Hz (i.e., 18 kbps) to 0.5 b/s/Hz (i.e., 90 kbps). Hence, in each frame of 5 ms duration, the average arrival rate to the transmission buffer varies from 90 to 450 bits. This implies the maximum delay bound corresponding to $Q^{\max}$ is about 200 ms to 1 s, which is suitable for IIoT applications such as smart grids, or smart healthcare. With the fixed power allocation, we assume an equal power allocation $P_{S_k} = P_{R_n} = 100$ mW (i.e., 20 dBm).

B. Convergence of the Online Iterative Algorithm

We first demonstrate the convergence of the online iterative algorithm in Section III. We assume $\mu_{S_k}^{\min} = 0.3$, and $\zeta_{S_k} = 10^{-4}, k = 1, 2$. For the relays, $\mu_{R_n}^{\min} = 0.3, n = 1, 2$, and $\zeta_{R_1} = 10^{-1}$, and $\zeta_{R_2} = 10^{-7}$ corresponding to the loose and stringent delay constraints, respectively.

Consider configuration C2. The decreasing step-size sequence is taken as $\epsilon[t] = 1/t^{0.3}$. The convergence results are shown in Fig. 2. We can observe that the time-averaged effective rates converge after 5×10^4 iterations. As expected, the capacities approach the minimum rate requirements of 0.3 for the sources and relays at convergence, implying no redundant power is used to support unnecessarily larger rates than the minimum required rates. Although the iterative algorithm is sub-optimal at the beginning, it quickly converges to the optimal solution as expected.

C. Fixed Power Allocation

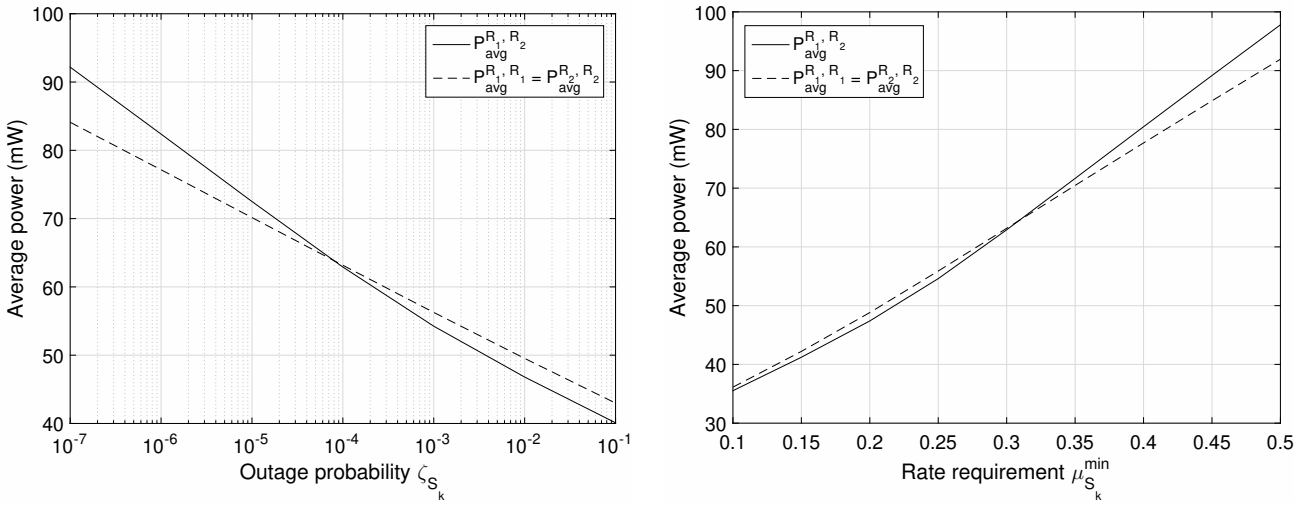
1) Effects of the delay constraints of the relays: We assume $\mu_{S_k}^{\min} = 0.3$, and $\zeta_{S_k} = 10^{-4}, k = 1, 2$. To study the effects of the delay constraints of the relays, we assume an equal value $\mu_{R_n}^{\min} = 0.3$ but different $\zeta_{R_n}, n = 1, 2$. Figure 3 shows the average powers of different configurations versus ζ_{R_2} assuming $\zeta_{R_1} = 10^{-4}$. It can be observed that any of the three configurations can be optimal depending on ζ_{R_2}. Particularly, for a very small ζ_{R_2}, it is optimal to select R_2 to assist both sources, which can be explained as follows. Given equal rate requirements of the relays, when R_2 has much more stringent delay constraint than R_1, the R_2–D link is scheduled much more frequently than the R_1–D link is in order to satisfy the delay constraint of R_2. As a result, both sources should be assisted by R_2 so that their data can be forwarded concurrently with that of relay R_2. Otherwise, the R_2–D link capacity would be under-utilized. For a similar reason, for a large ζ_{R_2}, it is optimal to select R_1 to assist both sources because R_1 now has a more stringent delay constraint than R_2 does. When both relays have almost similar delay constraint stringency, it is optimal that each relay assists the source closer to it. This is because the R_1–D link and the R_2–D link will be scheduled almost as often, and hence, each relay should just assist one source.

2) Effects of the minimum rate requirements of the relays: We assume an equal value for $\zeta_{R_n} = 10^{-3}, n = 1, 2$ but different values for $\mu_{R_n}^{\min}, n = 1, 2$.

Figure 4 displays the average powers of different network configurations versus $\mu_{R_2}^{\min}$ assuming $\mu_{R_1}^{\min} = 0.3$. As in the previous experiment, it can be seen that any of the three possible configurations can be optimal depending on $\mu_{R_2}^{\min}$. More specifically, for a small $\mu_{R_2}^{\min}$, selecting R_1 to assist both sources is optimal. When both relays have similar delay constraints, since R_1 has a much larger rate than R_2 does, the R_1–D link is scheduled more often. Consequently, both sources should be assisted by R_1 to utilize the capacity of the link R_1-D. Similarly, for a large $\mu_{R_2}^{\min}$, it is optimal to select R_2 to assist both sources. In other cases when the rate requirements of both relays are not much different, each relay should assist the source closer to it.

3) Effects of the delay constraints and rate requirements of the sources: The relays are assumed to have equal values of $\mu_{R_n}^{\min} = 0.3$ and $\zeta_{R_n} = 10^{-4}, n = 1, 2$.

Figure 5 show the average powers of different network configurations. In Fig. 5(a), we fix $\mu_{S_k}^{\min} = 0.3$ and vary ζ_{S_k}, and in Fig. 5(b), we fix $\zeta_{S_k} = 10^{-4}$ and vary $\mu_{S_k}^{\min}$. Note that for these settings, configuration C1 and C4 have similar power consumption. Again, depending on the rate requirements and



(a) versus ζ_{S_k} (b) versus $\mu_S^{\min}$

Fig. 5. Average powers for different configurations

delay constraints of the sources, either configuration can be optimal. For large rate requirements, and stringent delay constraints, it is better to select only one relay to assist both sources. This is because we can now utilize the capacity of the link from that relay to the destination to transmit the data of both sources. We can conclude that the optimal network configuration depends on the rate requirements and delay constraints of the sources and relays.

D. Adaptive Power Allocation

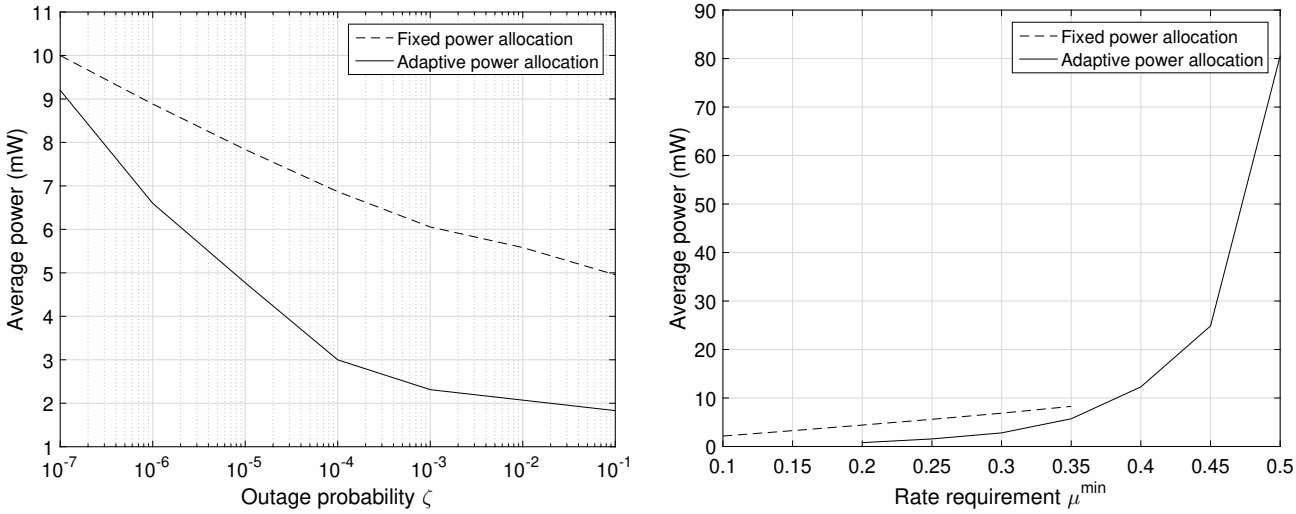
We will now demonstrate the benefits of the adaptive power allocation over the fixed power allocation schemes. The sources and relays are assumed to have the same rate requirements, i.e., $\mu_{S_k}^{\min} = \mu_{R_n}^{\min} = \mu^{\min}$, and delay constraints, i.e., $\zeta_{S_k} = \zeta_{R_n} = \zeta, k, n = 1, 2$. Moreover, for the fixed power allocation, we assume $P_{S_k} = P_{R_n} = 10$ mW. We show the average powers of both schemes when sources S_1 and S_2 are assisted by relays R_1 and R_2, respectively.

Figure 6 plot the average powers for fixed and adaptive power allocation schemes. While we fix $\mu^{\min} = 0.3$ and vary ζ in Fig. 6(a), we fix $\zeta = 10^{-4}$ and vary $\mu^{\min}$ in Fig. 6(b). It can be seen that the adaptive allocation achieves significant power gains over the fixed allocation. The gains are more profound under the looser delay constraints. This is because data transmissions of the sources and relays can be delayed for a longer time duration, and adaptive power allocation is efficient in exploiting the temporal fading diversity. On the other hand, under more stringent delay constraints, data transmissions need to take place more urgently, and the adaptive power allocation becomes less useful. Also, adaptive power allocation can achieve the rates beyond 0.35 (see Fig. 6(b)), which are impossible by the fixed power allocation.

Before concluding, it should be emphasized that the chosen parameters in this section is for illustrative purposes only to demonstrate the effects of the rate requirements and delay constraints on the power consumption and optimal routing configuration. In practice, depending on operating specifications, much lower power consumption can be used, which suits real-life IoT applications with low-power sensor nodes. Example scenarios include environmental monitoring applications with small data rate requirements, relaxed delay constraints, good channel conditions, which are typically true for IoT. Moreover, aspects of NB-IoT technology are not considered in the simulations. For example, re-transmission mechanism is not considered as reliable communications is implicitly assumed, for example, by using strong codes. Nevertheless, typical parameters in narrowband-IoT (NB-IoT) technology such as bandwidth of 180 kHz, frame duration of 5 ms etc. have been used, implying that with some suitable modifications, the simulation setting can be applied for small NB-IoT systems such as in home healthcare monitoring or manufacturing process monitoring with a wide range of rate requirements and delay constraints [25].

VI. Conclusions

We have studied the optimal resource allocation (i.e., joint link scheduling, rate, and power allocation) problems in dual-hop IoT networks. To provide delay QoS guarantees, we have imposed delay-outage constraints on the end-to-end sum queuing delays of the data flows of the nodes. The goal is to minimize the total power under the minimum rate requirements and delay constraints of the nodes. The proposed solution employs asymptotic delay analysis to compute the achieved delay exponents of the queue length tail distributions so that the delay-outage constraints are satisfied. After deriving the relation between the delay exponents and resource allocation variables, the solutions to the resulting constrained optimization problems are obtained using Lagrangian approach and convex optimization. We have also developed online algorithms based on stochastic approximation iterations when the fading statistics are *a-priori* unknown. Numerical studies have demonstrated the effects of delay constraints and rate requirements on the optimal power consumption and network configuration. Interesting future works include: 1) Extension of the presented design for general multi-hop networks and;



(a) versus ζ

(b) versus $\mu^{\min}$

Fig. 6. Average powers for fixed and adaptive power allocation schemes

2) Development of testbeds/prototypes for practical IIoT applications with experimental results to validate the theoretical results presented in this work. Also, it would be interesting to consider the non-cellular based protocols such asZigbee, 802.15.4, WirelessHART in the design and optimization, and model validation of resource allocation schemes.

REFERENCES

[1] J. Gubbi *et al.*, "Internet of Things (IoT): A Vision, Architectural Elements, and Future Directions," *Future Generation Computer Systems*, vol. 29, no. 7, pp. 1645–1660, 2013.

[2] E. Sisinni *et al.*, "Industrial Internet of Things: Challenges, Opportunities, and Directions," *IEEE Trans. Ind. Inform.*, vol. 18, no. 11, pp. 4724–4734, 2018.

[3] L. D. Xu *et al.*, "Internet of Things in Industries: A Survey", *IEEE Trans. Ind. Inform.*, vol. 10, no. 4, pp. 2233–2243, 2014.

[4] V. C. Gungor *et al.*, "Industrial Wireless Sensor Networks: Challenges, Design Principles, and Technical Approaches," *IEEE Trans. Ind. Inform.*, vol. 56, no. 10, pp. 4258–4265, 2009.

[5] P. Daugherty *et al.*, "Driving Unconventional Growth through the Industrial Internet of Things,". Accenture. Available online.

[6] R. Arshad *et al.*, "Green IoT: An Investigation on Energy Saving Practices for 2020 and Beyond," *IEEE Access*, vol. 5, pp. 15667–15681, 2017.

[7] S. Li *et al.*, "Energy-efficient Resource Allocation for Industrial Cyber-Physical IoT Systems in 5G Era," *IEEE Trans. Ind. Inform.*, vol. 14, o. 6, pp. 2618–2628, June 2018.

[8] T. Yamamoto *et al.*, "Multi-Hop Wireless Network for Industrial IoT," available online.

[9] S. Jeschke *et al.* (Eds.). Industrial Internet of Things. Springer. https://link.springer.com/book/10.1007%2F978-3-319-42559-7

[10] Z. Sheng *et al.*, "Recent Advances in Industrial Wireless Sensor Networks Toward Efficient Management in IoT," *IEEE Access*, vol. 3, pp. 622–637, 2015.

[11] C. Zhang *et al.*, "Throughput Optimization With Delay Guarantee for Massive Random Access of M2M Communications in Industrial IoT," *Internet of Things Journal,* vol. 6, no. 6, pp. 10077–10092, Dec. 2019.

[12] S. Bhandari *et al.*, "Latency Minimization in Wireless IoT Using Prioritized Channel Access and Data Aggregation," *IEEE GLOBECOM*, Dec. 2017.

[13] K. T. Phan *et al.*, "Power Allocation in Wireless Multi-user Relay Networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 5, pp. 2535–2545, 2009.

[14] S. Atapattu *et al.*, "Relay Selection and Performance Analysis in Multiple-user Networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 8, pp. 1–13, 2013.

[15] A. Ikhlef *et al.*, "Max-max Relay Selection for Relays with Buffers," *IEEE Trans. Wireless Commun.*, vol. 11, no. 3, pp. 1124–1135, 2012.

[16] I. Krikidis *et al.*, "Buffer-Aided Relay Selection for Cooperative Diversity Systems without Delay Constraints," *IEEE Trans. Wireless Commun.*, vol. 11, no. 5, pp. 1957–1967, 2012.

[17] N. Nomikos *et al.*, "A Buffer-aided Successive Opportunistic Relay Selection Scheme with Power Adaptation and Inter-Relay Interference Cancellation for Cooperative Diversity Systems," *IEEE Trans. Commun.*, vol. 63, no. 5, pp. 1623–1634, 2015.

[18] T. Islam *et al.*, "Multi-source Buffer-aided Relay Networks: Adaptive Rate Transmission," in *Proc. 2013 IEEE GLOBECOM*, Atlanta, USA.

[19] D. Wu *et al.*, "Effective Capacity: A Wireless Link Model for Support of Quality of Service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, 2003.

[20] L. Georgiadis, M. J. Neely, and L. Tassiulas. *Resource Allocation and Cross-Layer Control in Wireless Networks*. Foundations and Trends in Networking, vol. 1, no. 1, pp. 1–144, 2006.

[21] K. T. Phan *et al.*, "Optimal Resource Allocation for Buffer-Aided Relaying with Statistical QoS Constraint," *IEEE Trans. Commun.*, vol. 64, no. 3, pp. 959–972, 2016.

[22] C.-S Chang, "Stability, Queue Length, and Delay of Deterministic and Stochastic Queuing Networks," *IEEE Trans. Auto. Control*, vol. 39, no. 5, pp. 913–931, 1994.

[23] X. Wang *et al.* "Resource Allocation for Wireless Multiuser OFDM Networks," *IEEE Trans. Info. Theory*, vol. 57, no. 7, pp. 4359–4372, July 2011.

[24] A. Bhorkar *et al.*, "Power Optimal Opportunistic Scheduling," *IEEE GLOBECOM*, 2006.

[25] NB-IoT: Enabling New Business Opportunities. Huawei white paper. Available online: https://www.huawei.com/minisite/iot/img/nb_iot_whitepaper_en.pdf