Energy-Efficient Resource Allocation for Latency-Constrained Wireless Transmissions

Rong Chai and Shaowei Wang
Key Laboratory of Optoelectronic Devices and Systems with Extreme Performances of MOE and
School of Electronic Sciene and Engineering, Nanjing University, Nanjing 210023, China
Email: rongchai@smail.nju.edu.cn, wangsw@nju.edu.cn

2024 IEEE/CIC International Conference on Communications in China (ICCC) | 979-8-3503-7841-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICCC62479.2024.10682040

***Abstract*—In this paper, we propose an energy-efficient resource allocation strategy for wireless transmissions with latency guarantees. We establish a queue with a dynamic threshold structure at the transmitter to detect the occurrence of delay violation events. An energy-efficient resource allocation strategy based on Lyapunov optimization theory is proposed, which can effectively reduce average power consumption while adhering to latency constraints. Numerical results demonstrate that our method outperforms comparative strategies in various scenarios, effectively reducing average power consumption while ensuring over 99.9999% probability of meeting latency limits.**

***Index Terms*—Deterministic communications, energy-efficient resource allocation, latency-constrained wireless transmissions.**

I. Introduction

The rapid development of Internet of Things has led to a significant proliferation of wireless terminal devices and associated data demands [1]. Data transmission in time-sensitive applications such as autonomous driving, industrial control, and remote healthcare must adhere to strict latency constraints with high reliability [2]. 5G and beyond mobile networks are dediciated to providing determinstc communication services with end-to-end latency guarantees, attracting widespread attention from both academia and industry [3]–[5]. Achieving ultra-high reliability (surpassing 99.999%) for end-to-end latency assurance in wireless transmission systems is challenging due to channel stochasticity and limited resources, necessitating analysis and control of extreme events in delay distribution tail. Simultaneously, there is a concerted effort to advance energy-efficient communication, aiming to reduce system power consumption while ensuring latency constraints [6].

Regarding the existing literature, the vast majority considers mean delay as a primary metric for assessing system performance. In [7], researchers investigated the minimization of mean delay in single-link data transmission under Markovian arrivals. They modeled the sequential arrival data using a queue with an infinite buffer, formulating the optimization of mean delay under power constraints as a constrained Markov decision process. A cross-layer transmission strategy is proposed, integrating data arrivals, queue backlog, and channel conditions, with its optimality validated using the Lagrangian relaxation method. In [8], a framework utilizing parallel queues with finite buffers is employed to represent a time-sensitive wireless communication system consisting of a single server and multiple terminal devices. The authors propose a cross-layer scheduling scheme based on mean-field approximation theory to mitigate average data loss caused by violation of mean delay constraints.

Mean delay only reflects the first-order statistical characteristic, which is inadequate to fully characterize the entire delay distribution. Data transmission strategies solely focus on mean delay constraints are insufficient to control the occurrence of delay violation events, thus failing to meet the strict latency requirements of ultra-reliable and low-latency communication applications [9]. Due to the stochastic nature of traffic arrivals and channel conditions, ensuring deterministic delay bounds often leads to high costs [10]. Therefore, constraining the delay violation probability (DVP) has become a practical method to providing performance assurance for wireless transmission. In [11], a proficient inter-slice radio resource allocation scheme for mobile networks is presented, leveraging the network calculus model [12] to analyze system service latency and its variations, while also offering an upper bound for DVP. The latency performance bounds of the network calculus model are conservative, as it considers extreme scenarios with low occurrence probabilities, restricting full resource utilization. In [13], the authors investigate point-to-point communication systems equipped with automatic retransmission mechanisms. They employ a reinforcement learning agent optimized with proximal policy to dynamically control transmission power and bit rate. The authors statistically analyze the frequency of delay violation events and design penalty functions using heuristic methods, without providing a theoretical basis for latency guarantees. In [14], a transmission power allocation strategy for single-link communication is designed based on large deviation theory, featuring linear and higher decay exponent terms for DVP. Since their analysis is grounded on deterministic traffic, the conclusions may not apply to service scenarios with random data arrivals.

In this paper, we investigate resource allocation for wireless transmissions with latency constraints. We introduce a first-in-first-out queue at the transmitter and devise a dynamic threshold to detect the occurrence of delay violation events based on queue backlog. An energy-efficient resource allocation strategy based on Lyapunov optimization theory is proposed, which can effectively reduce average power consumption while adhering

This work was partially supported by the National Natural Science Foundation of China under Grants 61931023.

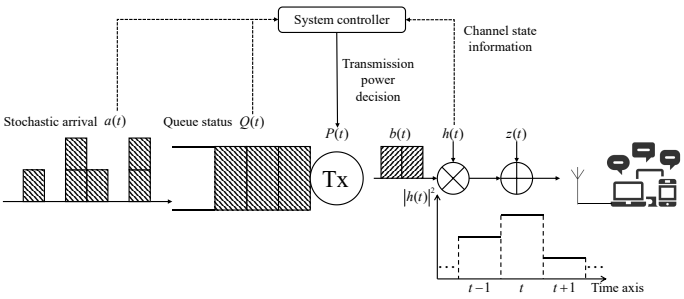

Fig. 1: System model.

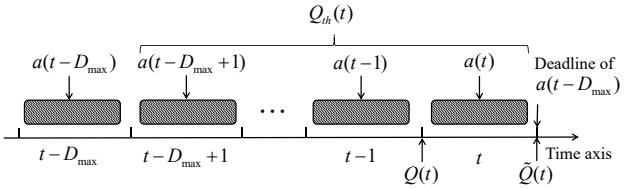

Fig. 2: Illustration of queue backlog and delay limit.

to DVP constraints. Numerical results demonstrate that our method outperforms comparison strategies by efficiently reducing system energy consumption and consistently meeting latency constraints with a probability exceeding 99.9999%.

II. System Model and Problem Formulation

A. System Model

Consider a point-to-point wireless transmission system, as depicted in Fig. 1, where the system controller dynamically adjusts power for data transmission to the receiver through a time-varying channel. We divide time into a series of discrete slots, represented by $t \in \{0, 1, 2, \cdots\}$, with each slot lasting for a duration of T seconds. The slot width T is sufficiently short to treat the channel coefficient as a constant within each slot. Let $h(t) \in \mathbb{C}$ denote the channel coefficient in slot t, which is assumed to be independent and identically distributed (i.i.d.) over slots. We assume that the system controller can accurately measure $h(t)$ using channel estimation techniques. According to *Shannon's formula*, with transmission power $P(t)$, the amount of data transmitted in slot t is determined as follows:

$$b(t) = BT \log_2 \left[1 + \frac{P(t) \left| h(t) \right|^2}{BN_0} \right], \tag{1}$$

where B represents the bandwidth and N_0 denotes the average power spectral density of additive noise $z(t)$.

Let $a(t)$ denote the arrival data at slot t. Without loss of generality, we assume that the stochastic arrival process is i.i.d. across all slots, with its mean denoted by λ. Due to the limited transmission rate, data arriving in each slot may not be immediately transmitted, potentially leading to data congestion within the system. The transmitter is equipped with sufficient memory space to store data awaiting transmission, represented by a queue with an infinite buffer operating on a first-in-first-out service order. Let $Q(t)$ denote the queue length at the beginning of slot t. By the end of this slot, the queue length will become:

$$\widetilde{Q}(t) = \max \left[Q(t) + a(t) - b(t), 0 \right]. \tag{2}$$

For time-sensitive applications, transmission delay is constrained by $D_{\max}$, representing the maximum number of slots that data is allowed to experience from entering the system to completing its transmission. Any data unable to meet this deadline will lose its utility and be discarded [13]. Assuming the transmitter and receiver are close enough to neglect radio propagation time, we approximate transmission delay as queuing delay. As shown in Fig. 2, all data that arrived at slot $t - D_{\max}$ must be fully transmitted by the end of slot t, otherwise, it exceeds the delay limit. In such cases, we refer to the occurrence as a delay violation event.

Let $Q_{th}(t) = \sum_{\tau=t-D_{\max}+1}^{t} a(\tau)$ represent the total arrival data volume from slot $t - D_{\max} + 1$ to t. We can determine whether transmission delay exceeds the limit by comparing $\widetilde{Q}(t)$ and $Q_{th}(t)$. If $\widetilde{Q}(t) > Q_{th}(t)$, a delay violation event occurs in slot t. The system controller will discard data in the buffer that exceeds its deadline. The queue length updates as follows:

$$Q(t+1) = \min \left[\widetilde{Q}(t), Q_{th}(t) \right]. \tag{3}$$

B. Problem Formulation

We aim to develop a power allocation strategy, which ensures that the probability of meeting latency constraint is not lower than $p_{th} \in [0, 1]$. Let $\epsilon_{th} = 1 - p_{th}$ denote the DVP constraint in the system. To ensure this constraint, the following equation should hold:

$$\lim_{t\to\infty} \frac{1}{t - D_{\max} + 1} \sum_{\tau=D_{\max}}^{t} \mathbb{E} \left\{ \mathbb{I} \left[\widetilde{Q}(\tau) > Q_{th}(\tau) \right] \right\} \leq \epsilon_{th}, \tag{4}$$

where $\mathbb{I}(\cdot)$ represents the indicator function, which takes the value of 1 when a delay violation event occurs.

Ensuring that the transmission rate is not lower than the arrival rate is essential for queue stability [15], leading to the following constraint:

$$\lim_{t\to\infty} \frac{1}{t+1} \sum_{\tau=0}^{t} \mathbb{E} \left[b(\tau) \right] \geq \lambda. \tag{5}$$

The system controller observes the arrival data quantity $a(t)$, queue backlog $Q(t)$, and channel coefficient $h(t)$ at each slot. It then makes a transmission power decision $P(t)$ based on established strategies. Our objective is to devise a power allocation strategy that, while satisfying the DVP constraint (4) and the transmission rate constraint (5), minimizes the average power consumption over the long term. We formulate

this online power decision problem as follows:

$$\begin{aligned}\min_{P(t)} \quad & \lim_{t\to\infty}\frac{1}{t+1}\sum_{\tau=0}^{t}\mathbb{E}\left[P(\tau)\right],\\ \text{s.t.} \quad & C_1:\lim_{t\to\infty}\frac{1}{t-D_{\max}+1}\sum_{\tau=D_{\max}}^{t}\mathbb{E}\left\{\mathbb{I}\left[\widetilde{Q}(\tau)>Q_{th}(\tau)\right]\right\}\\ & \le\epsilon_{th},\\ & C_2:\lim_{t\to\infty}\frac{1}{t+1}\sum_{\tau=0}^{t}\mathbb{E}\left[b(\tau)\right]\ge\lambda,\\ & C_3:0\le P(t)\le P_{\max},\quad\forall t,\end{aligned}\tag{6}$$

where $P_{\max}$ is the maximum transmission power available at the transmitter.

III. Energy-Efficient Resource Allocation with Latency Guarantee

We introduce two virtual queues $Z(t)$ and $H(t)$, corresponding to constraints C_1 and C_2 in (6), respectively. The virtual queues update as follows:

$$\begin{aligned}Z(t+1)&=\max\left\{Z(t)+\mathbb{I}\left[\widetilde{Q}(t)>Q_{th}(t)\right]-\epsilon_{th},0\right\},\\ H(t+1)&=\max\left[H(t)+\lambda-b(t),0\right].\end{aligned}\tag{7}$$

Here, $Z(0)=H(0)=0$. According to Lyapunov optimization theory, as long as $Z(t)$ and $H(t)$ satisfy $\lim_{t\to\infty}\frac{\mathbb{E}[Z(t)]}{t}=\lim_{t\to\infty}\frac{\mathbb{E}[H(t)]}{t}=0$, they are considered mean rate stable, and the corresponding constraints C_1 and C_2 in (6) can also be met [16].

Let $\mathbf{\Theta}(t)=[Z(t),H(t)]$. We define the Lyapunov function as follows:

$$L\left[\mathbf{\Theta}(t)\right]=\frac{1}{2}Z(t)^2+\frac{1}{2}H(t)^2.\tag{8}$$

We define the temporal difference of function (8) as the Lyapunov drift, denoted by $\Delta L(t)=L\left[\mathbf{\Theta}(t+1)\right]-L\left[\mathbf{\Theta}(t)\right]$, indicating the growth rate of the Lyapunov function. Given that function (8) increases with $Z(t)$ and $H(t)$, minimizing this drift at each slot ensures that function (8) remains stable or grows slowly over the long term, thereby maintaining the stability of virtual queues.

We define the following drift-plus-penalty structure, representing a trade-off between minimizing power consumption and constraint violation:

$$\mu\mathbb{E}\left[P(t)|\mathbf{\Theta}(t-1)\right]+\Delta L(t).\tag{9}$$

The parameter $\mu>0$ is denoted as the penalty factor. We can derive an upper bound for the drift-plus-penalty structure using $\max(x,0)^2\le x^2$:

$$\begin{aligned}&\mu\mathbb{E}\left[P(t)|\mathbf{\Theta}(t-1)\right]+\Delta L(t)\le C+\mu\mathbb{E}\left[P(t)|\mathbf{\Theta}(t-1)\right]\\ &\quad+H(t)\mathbb{E}\left[\lambda-b(t)|\mathbf{\Theta}(t-1)\right]\\ &\quad+Z(t)\mathbb{E}\left\{\mathbb{I}\left[\widetilde{Q}(t)>Q_{th}(t)\right]-\epsilon_{th}|\mathbf{\Theta}(t-1)\right\},\end{aligned}\tag{10}$$

where C is a constant satisfying:

$$\begin{aligned}C\ge&\frac{1}{2}\mathbb{E}\left[\lambda-b(t)|\mathbf{\Theta}(t-1)\right]^2+\\ &\frac{1}{2}\mathbb{E}\left\{\mathbb{I}\left[\widetilde{Q}(t)>Q_{th}(t)\right]-\epsilon_{th}|\mathbf{\Theta}(t-1)\right\}^2,\end{aligned}\quad\forall t.\tag{11}$$

Minimizing the upper bound of the drift-plus-penalty structure at each slot allows us to obtain a feasible solution for the original problem (6) while ensuring the stability of virtual queues and reducing power consumption [16]. We obtain the transmission power at each slot by solving the following subproblem:

$$\begin{aligned}\min_{P(t)} \quad & \mu P(t)+H(t)\left[\lambda-b(t)\right]\\ &+Z(t)\left\{\mathbb{I}\left[\widetilde{Q}(t)>Q_{th}(t)\right]-\epsilon_{th}\right\},\\ \text{s.t.} \quad & 0\le P(t)\le P_{\max}.\end{aligned}\tag{12}$$

Due to the presence of indicator function, problem (12) is non-convex and exhibits discontinuities, making it challenging to solve directly. To obtain its optimal solution, we consider the following variant:

$$\begin{aligned}\min_{P(t)} \quad & \mu P(t)+\\ &H(t)\left\{\lambda-BT\log_2\left[1+\frac{P(t)\left|h(t)\right|^2}{N_0B}\right]\right\},\\ \text{s.t.} \quad & P_{th,1}\le P(t)\le P_{th,2},\end{aligned}\tag{13}$$

where $P_{th,1}$ and $P_{th,2}$ belong to $\left(-\frac{N_0B}{|h(t)|^2},+\infty\right)$, with $P_{th,2}\ge P_{th,1}$. Problem (13) is convex, with its optimal solution derived analytically as follows:

$$P^*(t)=\begin{cases}P_{th,1}, & H(t)\le\frac{N_0\mu\ln 2}{T|h(t)|^2}+\frac{\mu P_{th,1}\ln 2}{BT},\\ \frac{BTH(t)}{\mu\ln 2}-\frac{N_0B}{|h(t)|^2}, & \frac{N_0\mu\ln 2}{T|h(t)|^2}+\frac{\mu P_{th,1}\ln 2}{BT}<H(t)\\ & <\frac{N_0\mu\ln 2}{T|h(t)|^2}+\frac{\mu P_{th,2}\ln 2}{BT},\\ P_{th,2}, & H(t)\ge\frac{N_0\mu\ln 2}{T|h(t)|^2}+\frac{\mu P_{th,2}\ln 2}{BT}.\end{cases}\tag{14}$$

Let $\Delta(t)=Q(t)+a(t)-Q_{th}(t)$ represent the minimum data transmission volume required at slot t to prevent a delay violation event. When $\Delta(t)<0$ or $\Delta(t)>BT\log_2\left[1+\frac{P_{\max}|h(t)|^2}{N_0B}\right]$, $Z(t)\left\{\mathbb{I}\left[\widetilde{Q}(t)>Q_{th}(t)\right]-\epsilon_{th}\right\}$ remains unchanged with $P(t)\in[0,P_{\max}]$. In such cases, the optimal solution to subproblem (12) is also the optimal solution to problem (13). We set $P_{th,1}=0$ and $P_{th,2}=P_{\max}$, substituting them into (14) to obtain the optimal transmission power.

If $0<\Delta(T)<BT\log_2\left[1+\frac{P_{\max}|h(t)|^2}{N_0B}\right]$, we first calculate the minimum transmission power required to send all of $\Delta(t)$:

$$P_1=\frac{N_0B\left[2^{\frac{\Delta(t)}{BT}}-1\right]}{|h(t)|^2}.\tag{15}$$

If $P(t)$ is not less than P_1, $\Delta(t)$ will be fully transmitted before its deadline. Thus, there will be no delay violation event in slot t.

Consider the following two policies: the first one avoids a delay violation event by selecting the transmission power in the interval $[P_1, P_{\max}]$, while the second one permits such an event, choosing the transmission power in the interval $[0, P_1)$. We denote the minimum values of the objective functions of (12) and (13) obtained by the first policy as V_1 and V_1', respectively. For the second policy, they are denoted as V_2 and V_2', respectively.

When $H(t) \geq \frac{N_0 \mu \ln 2}{T|h(t)|^2} + \frac{\mu P_1 \ln 2}{BT}$, the optimal solution to problem (13) lies within the interval $[P_1, P_{\max}]$, indicating $V_1' < V_2'$. Given $V_1 = V_1' - \epsilon_{th} Z(t)$ and $V_2 = V_2' + (1 - \epsilon_{th})Z(t)$, with $\epsilon_{th} \in [0, 1]$ and $Z(t) \geq 0$, it follows that $V_1 < V_2$. Consequently, the optimal solution of the subproblem (12) must reside within the interval $[P_1, P_{\max}]$. We set $P_{th,1} = P_1$ and $P_{th,2} = P_{\max}$, substituting them into (14) to obtain the optimal transmission power.

When $H(t) < \frac{N_0 \mu \ln 2}{T|h(t)|^2} + \frac{\mu P_1 \ln 2}{BT}$, the optimal solution to problem (13) lies within the interval $[0, P_1)$. We need to compare V_1 and V_2 to make power allocation decisions. For the first policy, we compute V_1 with transmission power P_1. For the second policy, we substitute $P_{th,1} = 0$ and $P_{th,2} = P_1$ into (14) to obtain the transmission power P_2, then calculate V_2. If $V_1 \leq V_2$, the optimal transmission power is P_1; otherwise, it is P_2.

Algorithm 1 outlines the details of the resource allocation strategy proposed by us.

Algorithm 1: Energy-Efficient Resource Allocation Strategy with Latency Guarantee

Input: Given arrival rate λ, delay constraint $D_{\max}$, and DVP constraint ϵ_{th}. Set penalty factor μ and maximum iterations $T_{\max}$.

$Q(0) \leftarrow 0, Z(0) \leftarrow 0, H(0) \leftarrow 0, t \leftarrow 0$;
while $t \leq T_{\max}$ **do**
  $Q_{th}(t) \leftarrow \sum_{\tau=t-D_{\max}+1}^{t} a(\tau)$;
  $\Delta(t) \leftarrow Q(t) + a(t) - Q_{th}(t)$;
  if $\Delta(t) < 0$ *or* $\Delta(t) > BT \log_2 \left[1 + \frac{P_{\max}|h(t)|^2}{N_0 B}\right]$ **then**
    $P_{th,1} \leftarrow 0, P_{th,2} \leftarrow P_{\max}$;
    Obtain transmission power $P^*(t)$ by (14);
  end
  else
    Obtain transmission power P_1 by (15);
    $H_1 \leftarrow \frac{N_0 \mu \ln 2}{T|h(t)|^2} + \frac{\mu P_1 \ln 2}{BT}$;
    if $H(t) \geq H_1$ **then**
      $P_{th,1} \leftarrow P_1, P_{th,2} \leftarrow P_{\max}$;
      Obtain transmission power $P^*(t)$ by (14);
    end
    else
      $P_{th,1} \leftarrow 0, P_{th,2} \leftarrow P_1$;
      Obtain transmission power P_2 by (14);
      Calculate objective function values V_1 and V_2 corresponding to transmission power P_1 and P_2 by (12);
      if $V_1 \leq V_2$ **then**
        $P^*(t) \leftarrow P_1$;
      end
      else
        $P^*(t) \leftarrow P_2$;
      end
    end
  end
  Calculate data transmission volume $b(t)$ with transmission power $P^*(t)$ by (1);
  Update all queues $Q(t), Z(t), H(t)$ by (3) and (7);
  $t \leftarrow t + 1$;
end

IV. Numerical Results

We assume that the arrival data follows a Poisson distribution and the channel experiences Rayleigh fading, with the mean of $|h(t)|^2$ set to 1. The values of some parameters are listed below: $B = 150\text{KHz}$, $T = 1\text{ms}$, $N_0 = -30\text{dBm/Hz}$, and $P_{\max} = 40\text{dBm}$. We run all simulations over 100 million slots.

Fig. 3 illustrates how the penalty factor μ controls the trade-off between constraint violation and power conservation. We set $\lambda = 100$, $D_{\max} = 10$, and $p_{th} = 99.9999\%$, performing equilogarithmically spaced sampling to obtain 15 penalty factor values between 1.2×10^5 and 2×10^5. It is observed that the frequency of delay violation events increases with μ, while average power consumption decreases. We substitute frequency for probability to estimate the probability mass function of delay. Fig. 4 illustrates the probability mass functions corresponding to different penalty factor settings, demonstrating that smaller μ are advantageous in attenuating the tail of the delay distribution, thereby reducing DVP. This observation remains consistent across various scenarios with different arrival rates and constraints. As the proportion of power consumption in the objective function of (12) increases with μ, our strategy prioritizes the reduction of transmission power. Some data cannot be transmitted before the deadline, consequently leading to an increase in DVP. During system deployment, the penalty factor can be dynamically adjusted by monitoring the frequency of delay violations over time. We can employ binary search to find the optimal penalty factor that precisely satisfies the DVP constraint while minimizing average power consumption, as indicated by μ^* in Fig. 3. In the remaining simulations of this section, the penalty factors have all been optimized using binary search.

We compare our method with three other resource allocation strategies: even load, resource redundancy, and effective capacity. Referring to Policy 1 as outlined in [17], the even load strategy evenly distributes data transmission tasks across all slots leading up to their respective deadlines. If the power required to transmit the expected amount of data exceeds the maximum power of the transmitter, the data is transmitted at $P_{\max}$. In [15], traffic intensity is defined as $\rho = \frac{\lambda}{\overline{b}}$, where $\overline{b}$ denotes the average transmission rate. The resource redundancy strategy employs a fixed transmission power P to

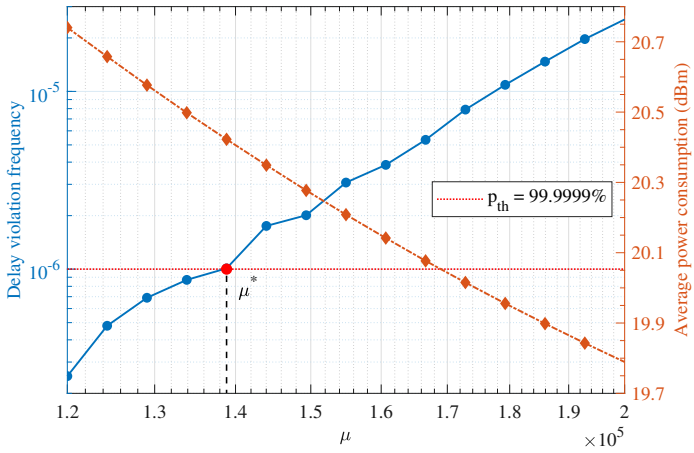

Fig. 3: Delay violation frequency and average power consumption with different penalty factor settings.

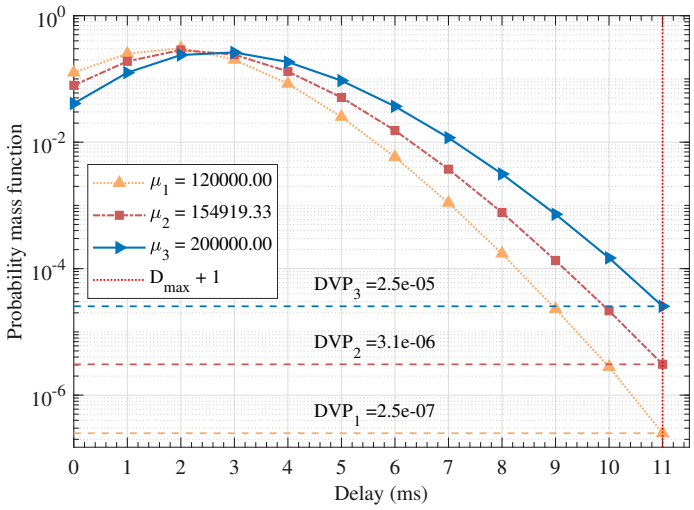

Fig. 4: Estimated probability mass functions with different penalty factor settings.

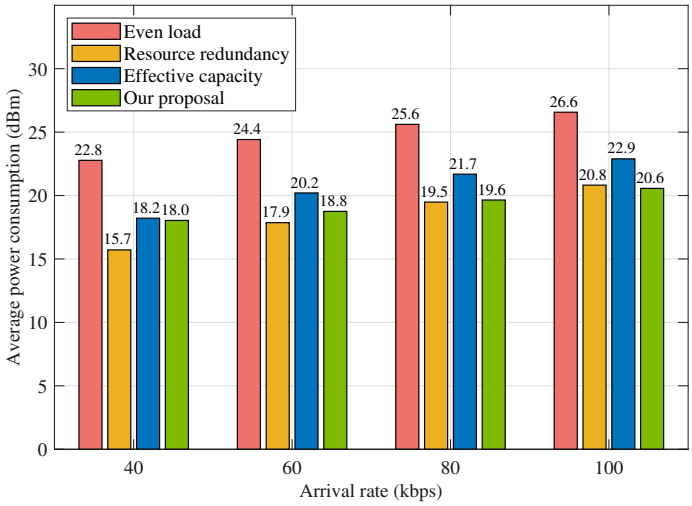

Fig. 5: Average power consumption against different arrival rates.

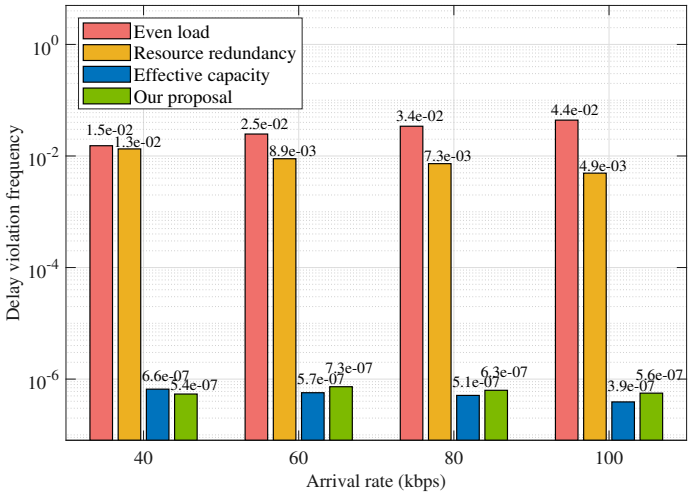

Fig. 6: Delay violation frequency against different arrival rates.

ensure $\rho < 1$. Here, we set $\rho = 0.9$ and numerically solve for the required constant transmission power. Effective capacity is derived from large deviation theory, which can provide statistical performance guarantee through static transmission strategy. In [18], an upper bound of DVP is provided for fixed transmission power based on effective capacity theory. The effective capacity strategy sets $\theta = -\frac{\ln \epsilon_{th}}{\lambda D_{\max}}$ and utilizes equation (13) from [18] to determine the necessary constant transmit power.

Fig. 5 and 6 show the average power consumption and delay violation frequency of various strategies under different arrival rates, respectively. We set the delay limit at $D_{\max} = 10$. Compared to the even load and resource redundancy strategies, the effective capacity strategy and our method consistently ensure compliance with the DVP constraint, achieving on-time data delivery with a probability exceeding $p_{th} = 99.9999\%$. This indicates that methods focusing on analyzing and constraining the delay distribution and its tail outperform strategies that solely focus on mean delay and provide redundant resources. In addition, our method demonstrates lower average power consumption compared to the effective capacity strategy. The simulation results confirm that our strategy can achieve data transmission with relatively low system power consumption while still meeting the DVP constraint, even under conditions of densely arriving data.

Fig. 7 and 8 depict the average power consumption and delay violation frequency of all strategies with different delay limits, respectively. We set the arrival rate at $\lambda = 100$. For our method, it is evident that with increasing $D_{\max}$, there is a decrease in average power consumption. This is attributed to the relaxation of the latency constraints, which enables our method to distribute data transmission tasks across different slots. Consequently, more data can be transmitted at lower power, particularly when the channel quality is favorable. Compared to other strategies, our method consistently achieves minimum average power consumption while meeting the DVP

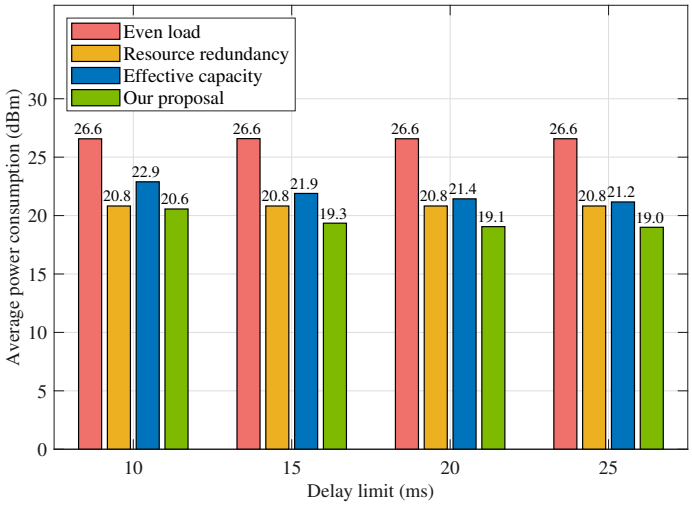

Fig. 7: Average power consumption against different delay limits.

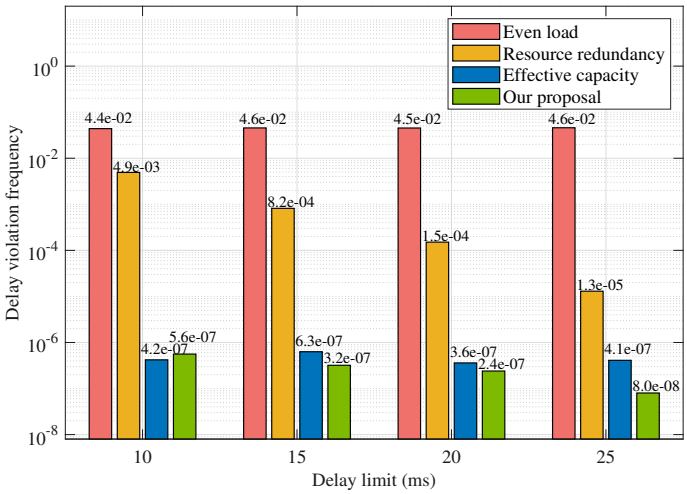

Fig. 8: Delay violation frequency against different delay limits.

constraints, thus confirming its effectiveness and robustness.

V. Conclusion

In this paper, we present an energy-efficient resource allocation strategy for latency-constrained wireless transmissions. We establish a queue with a dynamic threshold structure at the transmitter, evaluating the occurrence of delay violation events by checking whether the queue backlog exceeds the limit. Two virtual queues are introduced, with their stability maintained to ensure the satisfaction of corresponding constraints. We define the Lyapunov function and its drift, making transmission power decisions by minimizing the upper bound of the drift-plus-penalty structure at each slot. Simulation results across various scenarios indicate that our approach outperforms comparative strategies, achieving energy-efficient data transmission while ensuring a probability exceeding 99.9999% of meeting latency constraints. Our method offers insights for guiding resource allocation in practical wireless transmission systems.

References

[1] D. C. Nguyen *et al.*, "6G Internet of Things: A comprehensive survey," *IEEE Internet Things J.*, vol. 9, no. 1, pp. 359–383, Jan. 2022.

[2] K. B. Letaief *et al.*, "Edge artificial intelligence for 6G: Vision, enabling technologies, and applications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 5–36, Jan. 2022.

[3] T. Wang and S. Wang, "Inter-slice radio resource allocation: An online convex optimization approach," *IEEE Wirel. Commun.*, vol. 28, no. 5, pp. 171–177, Oct. 2021.

[4] S. Liu, T. Wang and S. Wang, "Hardware impairment estimation in NB-IoT: A parallel multitask learning method," *IEEE Internet Things J.*, vol. 10, no. 8, pp. 6859–6869, Apr. 2023.

[5] W. Yang *et al.*, "Semantic communications for future internet: Fundamentals, applications, and challenges," *IEEE Commun. Surv. Tutorials*, vol. 25, no. 1, pp. 213–250, 1th Quart. 2023.

[6] S. Buzzi *et al.*, "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.

[7] X. Zhao *et al.*, "Delay-optimal and energy-efficient communications with Markovian arrivals," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1508–1523, Mar. 2020.

[8] C. Li, W. Chen and K. B. Letaief, "Meeting hard delay constraint in massive access: A mean-field approach," *IEEE Trans. Wireless Commun.*, vol. 23, no. 4, pp. 2961–2977, Aug. 2023.

[9] M. Bennis, M. Debbah and H. V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale," *Proc. IEEE*, vol. 106, no. 10, pp. 1834–1853, Oct. 2018.

[10] L. Cheng *et al.*, "Adaptive forwarding with probabilistic delay guarantee in low-duty-cycle WSNs," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4775–4792, Jul. 2020.

[11] J. Zhu and S. Wang, "Delay-guaranteed resource allocation for deterministic communications: An efficient stochastic network calculus method," in *Proc. IEEE GLOBECOM'23*, Kuala Lumpur, Malaysia, Dec. 2023.

[12] A. Bouillard, M. Boyer and E. L. Corronc, *Deterministic Network Calculus: From Theory to Practical Implementation*. Hoboken, NJ, USA: John Wiley & Sons, 2018.

[13] H. Peng *et al.*, "Power and rate adaptation for URLLC with statistical channel knowledge and HARQ," *IEEE Wireless Commun. Lett.*, vol. 12, no. 12, pp. 2148–2152, Dec. 2023.

[14] L. Li, W. Chen and K. B. Letaief, "Simple bounds on delay-constrained capacity and delay-violation probability of joint queue and channel-aware wireless transmissions," *IEEE Trans. Wireless Commun.*, vol. 22, no. 4, pp. 2744–2759, Apr. 2023.

[15] I. Adan and J. Resing, *Queueing Theory*. Eindhoven University of Technology, 2002.

[16] M. J. Neely, *Stochastic Network Optimization with Application to Communication and Queueing Systems*. Cham, Switzerland: Springer, 2010.

[17] L. Huang, L. Li and W. Chen, "Diversity enabled wireless transmissions with random arrivals and hard delay constraints," in *Proc. IEEE ICC'23*, Rome, Italy, May 2023.

[18] J. Choi, "An effective capacity-based approach to multi-channel low-latency wireless communications," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2476–2486, Mar. 2019.