Drift Plus Optimistic Penalty – A Learning Framework for Stochastic Network Optimization

Sathwik Chadaga and Eytan Modiano
Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA

***Abstract*—We consider the problem of joint routing and scheduling in queueing networks, where the edge transmission costs are unknown. At each time-slot, the network controller receives noisy observations of transmission costs only for those edges it picks for transmission. The network controller's objective is to take routing and scheduling decisions so that the total expected cost is minimized. This problem exhibits an exploration-exploitation trade-off, however, previous bandit-style solutions cannot be directly applied to this problem due to the queueing dynamics. In order to ensure network stability, the network controller needs to optimize throughput and cost simultaneously. We show that the best achievable cost is lower bounded by the solution to a static optimization problem, and develop a network control policy using techniques from Lyapunov drift-plus-penalty optimization and multi-arm bandits. We show that the policy achieves a sub-linear regret of order $O(T^{2/3})$, as compared to the best policy that has complete knowledge of arrivals and costs. Finally, we evaluate the proposed policy using simulations and show that its regret is indeed sub-linear.**

***Index Terms*—Optimal network control, online shortest path routing, stochastic bandits, Lyapunov optimization.**

I. Introduction

Stochastic network optimization refers to the problem of making routing and scheduling decisions in network systems to optimize given objectives such as cost, utility, and throughput. It is fundamental in designing good network resource allocation strategies and has applications in many fields such as communications [1], [2], cloud computing [3], [9], and content delivery [4], [10]. Early studies on network optimization [1]–[3] consider a static version of this problem, where traffic rates are modeled as static flows. However, practical network systems are not static and require queue management due to the stochastic nature of traffic. Therefore, their control involves maintaining high throughput to ensure low queueing delays.

In [5], [6], throughput optimal policies are designed for power constrained wireless systems with known arrival rates. When arrival statistics are unknown, the Lyapunov optimization technique [7] is commonly used to develop throughput-optimal policies. A policy that jointly optimizes energy and throughput is proposed in [8]. It uses the Lyapunov drift-plus-penalty minimization technique to demonstrate a trade-off between throughput and energy consumption. This technique has since been used to design throughput-optimal and minimum-cost policies in many other fields. For example, Lyapunov optimization has been used to develop cost effective battery management schemes in data centers [9] and content distribution strategies in cloud infrastructures [10]. A similar power control strategy for wireless networks with batteries is studied in [11]. Finally, [12] combines Lyapunov minimization and shortest-path routing to design a minimum hop routing scheme. All of the above works assume that the transmission costs are known in advance.

In applications where costs represent quantities such as power consumption [13], operational cost [14], server load [15], or quality of service [16], costs may initially be unknown and need to be learned through feedback. Moreover, the costs are revealed only partially depending on the choice of action. This is addressed by the multi-arm bandit techniques [17] where, one has to pick the best arm among multiple arms each associated with an unknown cost. Whenever an arm is chosen, a noisy observation of its cost is received. Hence, this exhibits an exploration-exploitation trade-off. A widely used strategy to deal with this trade-off is the upper-confidence-bound strategy [18], [19] where, arms are chosen based on optimistically biased cost estimates, which allow exploration by lowering the costs of under-explored arms.

Routing in networks with unknown costs can also be formulated as a multi-arm bandit problem with each route acting as an arm. However, this method is not scalable as the number of routes (arms) grows exponentially with network size. Alternatively, one can exploit the fact that paths share edges in a network and hence path costs are dependent. Consequently, an efficient exploration basis called barycentric spanner that can be computed in polynomial time is proposed in [20]. Using this, the confidence-bound style exploration is extended to shortest path routing in [21]–[23].

These bandit-type solutions cannot be extended to queueing networks directly. In the traditional bandits problem, the best arm with the least underlying cost is optimal. The solution involves exploring enough arms and converging onto the best arm quickly. However, in queueing networks, the arrival rate is often larger than the capacity of a single path. Hence, minimizing the cost alone will cause queue backlogs to build up in certain low cost paths, which reduces network throughput and leads to instability. It is important that the network controller optimizes both cost and throughput simultaneously.

The problem of optimal routing and scheduling in a single-hop queueing system with unknown costs is considered in [24] where, a priority scheme based on queue lengths and upper-confidence-bound cost estimates is designed. Despite having a logarithmic regret, this policy is limited to single-hop systems. In multi-hop networks, the packets routed through

This work was supported by NSF grants CNS-2148183 and CNS-2106268.

one path will affect queue backlogs of other overlapping paths. Hence, extending this work to multi-hop networks is non-trivial. Further, [25] addresses multi-hop routing through traffic splitting at the source, by using an optimistically estimated version of a traffic splitting metric from [12]. However, the regret is defined as a function of only this traffic splitting metric and not as a function of the total cost. Hence, this does not address the total cost minimization problem.

In this paper, we consider the problem of optimal routing and scheduling in stochastic queueing networks. We assume that the arrival rates and edge transmission costs are unknown. Instead, the control policy has to make transmission decisions based only on the queue backlogs and past cost observations. We seek to design a policy with sub-linear regret, where the regret is defined as the gap between policy's expected cost and the cost of an optimal policy with complete knowledge of arrivals and costs. We summarize our contributions below.

- We define a novel cost metric in terms of both transmission costs and queue backlogs that allows us to jointly optimize cost and throughput. We show that the best achievable cost of any policy is lower bounded by the solution to a static optimization problem.
- We develop a network control policy by combining ideas from Lyapunov drift-plus-penalty minimization technique and the upper-confidence-bound algorithm. Further, we show that the proposed policy has a sub-linear regret of order $O(T^{2/3}+\sqrt{T}\log T)$, where T is the time horizon.
- We evaluate the proposed policy using simulations and show that it indeed has a sub-linear regret. We also simulate an oracle policy that knows the underlying costs exactly and show that the proposed policy's backlog and cost performance approaches the oracle's performance.

The rest of the paper is organized as follows. We describe our model and formulate the network control problem in Section II. We derive a lower bound on the best achievable cost in Section III. We discuss the proposed Drift Plus Optimistic Penalty policy in Section IV and analyze the policy's regret in Section V. Finally, we present the simulation results in Section VI and conclusions in Section VII.

II. Problem Formulation

a) Network and Traffic Model: We consider a multi-hop network $G=(\mathcal{N},\mathcal{E})$, where $\mathcal{N}$ is the set of nodes and $\mathcal{E}$ is the set of directed edges. We denote the source and destination nodes by s and d respectively[1]. The network operates at discrete time-slots $t=1,2,...,T$, where T is the given *time horizon*. At each time t, there are $a(t)$ packets that arrive at the source. The arrivals $a(t)$'s are independent random variables with an unknown *arrival rate* $\lambda:=\mathbb{E}[a(t)]$ packets/slot. The network maintains a first-in-first-out queue to buffer incoming packets at every node. We denote by $Q_i^\pi(t)$ the *queue backlog* at node $i\in\mathcal{N}$ and time t, and by $\{Q_i^\pi(t)\}_{i\in\mathcal{N}}$ the set of all queue backlogs at time t. A given control policy π observes the queue backlogs at each time-slot and decides the number of packets to be transmitted on each edge. We denote by $\mu_{ij}^\pi(t)$ the policy π's *planned transmission* in packets/slot on edge $(i,j)\in\mathcal{E}$ at time t, and by $\{\mu_{ij}^\pi(t)\}_{(i,j)\in\mathcal{E}}$ the set of all planned transmissions at time t. We denote the set of outgoing neighbors of any node $i\in\mathcal{N}$ by $\mathcal{N}_i:=\{j\in\mathcal{N}:(i,j)\in\mathcal{E}\}$. The queue backlog evolution can now be expressed as follows. At the source node s, $\forall t=1,2,...,T$,

$$Q_s^\pi(t+1)=\left[Q_s^\pi(t)-\sum_{j\in\mathcal{N}_s}\mu_{sj}^\pi(t)\right]^+ + a(t) \tag{1}$$

where, $[\cdot]^+=\max\{\cdot,0\}$. At other nodes, $\forall i\in\mathcal{N}$, $i\neq s,d$,

$$Q_i^\pi(t+1)\leq\left[Q_i^\pi(t)-\sum_{j\in\mathcal{N}_i}\mu_{ij}^\pi(t)\right]^+ + \sum_{j:i\in\mathcal{N}_j}\mu_{ji}^\pi(t). \tag{2}$$

Finally, we assume that packets exit the network immediately when they reach the destination node d. Hence $\forall t$,

$$Q_d^\pi(t)=0. \tag{3}$$

Notice that the queue evolution expression (2) is an inequality. This is because some queues may not have enough buffered packets to transmit the number of packets planned by the policy. As a result, the actual number of packets transmitted on some edges may be less than the planned number of transmissions. We denote by $\tilde{\mu}_{ij}^\pi(t)$ the *actual transmission* in packets/slot on edge $(i,j)\in\mathcal{E}$ at time t, and by $\{\tilde{\mu}_{ij}^\pi(t)\}_{(i,j)\in\mathcal{E}}$ the set of all actual transmissions at time t. Note that the actual transmissions are constrained by $\sum_{j\in\mathcal{N}_i}\tilde{\mu}_{ij}^\pi(t)\leq Q_i^\pi(t)$, $\forall i\in\mathcal{N}$ and $\tilde{\mu}_{ij}^\pi(t)\leq\mu_{ij}^\pi(t)$, $\forall(i,j)\in\mathcal{E}$.

b) Stability Region: We aim to keep the queue backlogs small so that our throughput is equal to the arrival rate. We keep track of queue backlogs using the notion of *mean rate stability* from [26]. A queueing network with backlogs $\{Q_i^\pi(t)\}_{i\in\mathcal{N}}$ is said to be mean rate stable under policy π if the queue backlogs satisfy $\lim_{T\to\infty}\frac{1}{T}\sum_{i\in\mathcal{N}}\mathbb{E}[Q_i^\pi(T)]=0$.

Further, each edge $(i,j)\in\mathcal{E}$ has a finite capacity denoted by μ_{ij}^{max}. A policy π is feasible if its transmission decisions satisfy $\forall t,\forall(i,j)\in\mathcal{E}$, $0\leq\mu_{ij}^\pi(t)\leq\mu_{ij}^{max}$. Let Π^* be the collection of all feasible control policies, including the policies with knowledge of future arrivals. We define the *stability region* $\Lambda(G)$ as the set of arrival rates for which there exists a policy $\pi\in\Pi^*$ that keeps the system stable. In [27], it was shown that the stability region $\Lambda(G)$ can be characterized as the set of all arrival rates λ for which there exists feasible flows $\{\mu_{ij}\}_{(i,j)\in\mathcal{E}}$ that satisfy the following conditions:

$$\left\{\begin{array}{c}\lambda\leq\sum_{j\in\mathcal{N}_s}\mu_{sj},\\ \sum_{j:i\in\mathcal{N}_j}\mu_{ji}\leq\sum_{j\in\mathcal{N}_i}\mu_{ij},\ \forall i\neq s,d,\\ 0\leq\mu_{ij}\leq\mu_{ij}^{max},\ \forall(i,j)\in\mathcal{E}.\end{array}\right\} \tag{4}$$

Formally, the stability region $\Lambda(G)$ is defined as

$$\Lambda(G):=\left\{\lambda\geq 0:\exists\ \{\mu_{ij}\}_{(i,j)\in\mathcal{E}}\text{ that satisfies (4)}\right\}.$$

Throughout this paper, we assume that the arrivals belong to the network stability region i.e. $\mathbb{E}[a(t)]\in\Lambda(G)$.

[1]We consider a single s-d pair for simplicity of exposition, however extension to multiple s-d pairs is straightforward.

c) Cost Structure: Each edge (i,j) has an unknown *transmission cost* of c_{ij} per packet. Hence, a given policy π incurs a total transmission cost of $\sum_{t=1}^{T}\sum_{(i,j)\in\mathcal{E}}\tilde{\mu}_{ij}^{\pi}(t)c_{ij}$. We assume that these costs are bounded by $0 \leq c_{ij} \leq C_{max}$. Further, minimizing the transmission cost alone may cause network instability. For example, a bad policy can attain zero cost by simply not transmitting any packets and letting the queues build up. To avoid this, we include the following queue backlog penalty. The policy incurs a *terminal backlog cost* $C_B \geq 0$ for each undelivered packet at the end of the time horizon T. Hence, the policy incurs a total backlog cost of $C_B\sum_{i\in\mathcal{N}}Q_i^{\pi}(T)$. Intuitively, we should pick a large terminal cost C_B so that the policy is encouraged to deliver packets to the destination. Otherwise, the policy can let queues build up and cause instability. This terminal cost can also be interpreted as the cost incurred to deliver the remaining packets at the end of time horizon T to the destination. We will discuss this in more detail in Section III. In summary, the total cost $C^{\pi}(T)$ incurred by a given control policy π is defined as follows:

$$C^{\pi}(T) := \sum_{t=1}^{T}\sum_{(i,j)\in\mathcal{E}}\tilde{\mu}_{ij}^{\pi}(t)c_{ij} + C_B\sum_{i\in\mathcal{N}}Q_i^{\pi}(T).$$

As actual transmissions $\{\tilde{\mu}_{ij}^{\pi}(t)\}_{(i,j)\in\mathcal{E}}$ are constrained both by edge capacities and queue states, working with them complicates the analysis. Hence, we simplify our analysis by expressing the total cost in terms of the planned transmissions $\{\mu_{ij}^{\pi}(t)\}_{(i,j)\in\mathcal{E}}$ instead. By definition, we have $\tilde{\mu}_{ij}^{\pi}(t) \leq \mu_{ij}^{\pi}(t)$. And since the costs are non-negative, we have

$$C^{\pi}(T) \leq \sum_{t=1}^{T}\sum_{(i,j)\in\mathcal{E}}\mu_{ij}^{\pi}(t)c_{ij} + C_B\sum_{i\in\mathcal{N}}Q_i^{\pi}(T). \tag{5}$$

This simplifies our analysis as the planned transmissions are only constrained by edge capacities and not by the queue sizes.

As the costs c_{ij}'s are initially unknown, the policy relies on the following semi-bandit feedback [19] it receives depending on its actions. At each t, the policy receives a *noisy observation* $\widetilde{c}_{ij}(t)$ for each edge on which it transmitted any packets,

$$\forall(i,j)\in\mathcal{E} : \mu_{ij}^{\pi}(t) > 0,\ \widetilde{c}_{ij}(t) := c_{ij} + \eta_{ij}(t)$$

where, $\eta_{ij}(t)$'s are zero mean σ-sub-Gaussian[2] random variables that are independent across time slots. As discussed before, the neighboring queues of some edges with $\mu_{ij}^{\pi}(t) > 0$ may not have enough packets to support the planned transmissions. On these edges, we assume that the policy can send null packets, no more than $\mu_{ij}^{\pi}(t)$, and still observe the costs. Note that these null packets do not contribute to the queue evolution but they do incur transmission costs. Also note that this assumption does not affect our analysis as we have expressed the cost as a function of planned transmissions in (5) rather than the actual transmissions. Denote by Π^* the collection of all policies, including those with knowledge of costs and future arrivals. We define the *regret* of policy π as

$$R^{\pi}(T) := \mathbb{E}[C^{\pi}(T)] - \inf_{\pi^*\in\Pi^*}\mathbb{E}[C^{\pi^*}(T)].$$

where, the expectation is taken over the randomness in arrivals and possibly in policy's actions. Note that, unlike the best policy π^*, the policies we consider will not have access to the cost values, future arrivals, and even the arrival rates. Let Π be the collection of admissible policies that do not know the costs $\{c_{ij}\}_{(i,j)\in\mathcal{E}}$, do not know the arrival rate λ, and make causal control decisions. We now state our objective as follows.

Objective: Find a policy $\pi \in \Pi$ that has sub-linear regret

$$\lim_{T\to\infty}\frac{R^{\pi}(T)}{T} = 0.$$

III. Static Lower Bound on the Optimal Cost

In this section, we obtain a lower bound on the optimal cost, using a static flow version of the problem. This bound will be useful in later sections when we perform regret analysis of our proposed policy. Consider the optimization problem $\mathcal{P}$.

$$\begin{aligned}\mathcal{P}:\ \min_{\mu}\ &\sum_{(i,j)\in\mathcal{E}}\mu_{ij}c_{ij},\\ \text{subject to }\ &\lambda \leq \sum_{j\in\mathcal{N}_s}\mu_{sj},\\ &\sum_{j:i\in\mathcal{N}_j}\mu_{ji} \leq \sum_{j\in\mathcal{N}_i}\mu_{ij},\ \forall i\in\mathcal{N}\setminus\{s,d\},\\ &0 \leq \mu_{ij} \leq \mu_{ij}^{max},\ \forall(i,j)\in\mathcal{E}.\end{aligned}$$

In the following theorem, we show that for a large enough terminal backlog cost C_B, the optimal value of $\mathcal{P}$ lower-bounds the best policy's cost. Let $\{\mu_{ij}^{stat}\}_{(i,j)\in\mathcal{E}}$ be the optimal solution to $\mathcal{P}$. Additionally, for a given network topology G, define by C_L the maximum total cost that a packet can incur when it travels from s to d using acyclic routes.

Theorem 1: (Static Lower Bound) For $C_B \geq C_L$, we have

$$\inf_{\pi^*\in\Pi^*}\mathbb{E}[C^{\pi^*}(T)] \geq T\sum_{(i,j)\in\mathcal{E}}\mu_{ij}^{stat}c_{ij}.$$

We prove Theorem 1 in Appendix A. Note that the theorem suggests that a good choice of C_B should be larger than C_L. Otherwise, a policy may be able to achieve lower costs by leaving more packets in the buffers at the end of time horizon, and fail to stabilize the network. Since our goal is to design stabilizing policies, we assume that $C_B > C_L$ for the rest of the paper. Further, note that even though the solution to the optimization problem $\mathcal{P}$ lower bounds the cost, a policy that chooses transmissions equal to μ_{ij}^{stat} at all time t is not optimal. This is because a portion of the arrived packets will remain in the queues due to the randomness in arrivals. Hence, including the queue backlog penalty, such a policy's total cost will be greater than $T\sum_{(i,j)\in\mathcal{E}}\mu_{ij}^{stat}c_{ij}$. Moreover, since the arrival rate and costs are unknown, it is impossible to calculate the solution to $\mathcal{P}$. We only use the solution to $\mathcal{P}$ as a bound to analyze our policy's regret and our policy will not be required to solve this optimization problem. Finally, as a

[2]A random variable X is σ-sub-Gaussian if $\mathbb{E}[e^{\zeta(X-\mathbb{E}[X])}] \leq e^{\frac{\sigma^2\zeta^2}{2}}, \forall\zeta$.

direct implication of Theorem 1, we can express the regret of any policy as follows.

Corollary 1: For $C_B \geq C_L$, we have

$$R^{\pi}(T) \leq \sum_{t=1}^{T} \sum_{ij \in \mathcal{E}} \mathbb{E}\left[\mu_{ij}^{\pi}(t) - \mu_{ij}^{stat}\right] c_{ij} + C_B \sum_{i \in \mathcal{N}} \mathbb{E}\left[Q_i^{\pi}(T)\right].$$

Proof of Corollary 1: The result directly follows from the definition of regret, the cost bound (5), and Theorem 1. □

It can be seen from Corollary 1 that any policy with sub-linear regret will also be mean rate stable. Hence, our objective of finding a policy that has sub-linear regret inherently ensures network stability. We discuss our proposed Drift Plus Optimistic Penalty policy in the next section.

IV. Drift Plus Optimistic Penalty Policy

We use the technique of drift-plus-penalty minimization from [8] to derive our policy. We first define the Lyapunov function under a given policy π at time t as

$$\Phi^{\pi}(t) := \frac{1}{2} \sum_{i \in \mathcal{N}} Q_i^{\pi}(t)^2.$$

We denote the Lyapunov drift at time t as $\Delta\Phi^{\pi}(t) := \Phi^{\pi}(t+1) - \Phi^{\pi}(t)$. Further, given queue backlogs, we define the Lyapunov drift-plus-penalty at time t as

$$L^{\pi}(t) := \mathbb{E}[\Delta\Phi^{\pi}(t) + \nu \sum_{(i,j) \in \mathcal{E}} \mu_{ij}^{\pi}(t) c_{ij} \mid \{Q_i^{\pi}(t)\}_{i \in \mathcal{N}}]$$

where, ν is a tuning parameter that we will use to tune the cost-backlog trade-off.

The idea behind drift-plus-penalty minimization technique is to greedily minimize $L^{\pi}(t)$ at each time-slot t. However, we cannot do this directly in our problem as the edge costs c_{ij} are unknown. Instead, we have to estimate the costs using past observations and make our decisions based on these estimates. Moreover, since we get observations for only the edges we select, we face an exploration-exploitation trade-off. We have to simultaneously exploit past observations and explore less observed edges to improve their cost estimates. This motivates us to use the idea of *optimism in face of uncertainty* from the multi-arm bandits literature [18], [19]. Recall that the observed cost values until time t is given by $\{\forall \tau < t,\ \forall (i,j) \in \mathcal{E} : \mu_{ij}^{\pi}(\tau) > 0,\ \widetilde{c}_{ij}(\tau)\}$. For each edge $(i,j) \in \mathcal{E}$, let $N_{ij}(t)$ be the number of observations received until time t. Denote by $\bar{c}_{ij}(t)$ the average of observations until time t,

$$\bar{c}_{ij}(t) = \frac{1}{N_{ij}(t)} \sum_{\tau=0}^{t-1} \widetilde{c}_{ij}(\tau) \mathbb{I}[\mu_{ij}^{\pi}(\tau) > 0]$$

where, $\mathbb{I}[\cdot]$ is the indicator function. Now we define the lower-confidence-bound estimate of the cost at time t as

$$\hat{c}_{ij}(t) := \bar{c}_{ij}(t) - \sqrt{\frac{\beta \log(t/\delta)}{N_{ij}(t)}} \tag{6}$$

where, $\delta \in (0,1)$ and $\beta \geq 0$ are tuning parameters. Note that this is an optimistic estimate of the costs, where the term $\sqrt{\beta \log(t/\delta)/N_{ij}(t)}$ adds a bias in favor of under-explored edges whose $N_{ij}(t)$ values are small. Now, as the true costs are unknown, we instead use these optimistic estimates in the drift-plus-penalty expression. Formally, we use the new drift-plus-optimistic-penalty expression $\hat{L}^{\pi}(t)$ defined as

$$\hat{L}^{\pi}(t) := \mathbb{E}_{|Q^{\pi}, \hat{c}}\left[\Delta\Phi^{\pi}(t) + \nu \sum_{(i,j) \in \mathcal{E}} \mu_{ij}^{\pi}(t) \hat{c}_{ij}(t)\right]$$

where, $\mathbb{E}_{|Q^{\pi}, \hat{c}}[\cdot] := \mathbb{E}[\cdot \mid \{Q_i^{\pi}(t)\}_{i \in \mathcal{N}}, \{\hat{c}_{ij}(t)\}_{(i,j) \in \mathcal{E}}]$ the conditional expectation given queue backlogs and cost estimates. From the queue evolution dynamics from Section II and from the fact that $([q-b]^+ + a)^2 \leq q^2 + b^2 + a^2 + 2q(a-b)$, we obtain the following bound on drift-plus-optimistic-penalty.

$$\hat{L}^{\pi}(t) \leq B + \lambda Q_s^{\pi}(t) + \mathbb{E}_{|Q^{\pi}, \hat{c}}\left[\sum_{(i,j) \in \mathcal{E}} \mu_{ij}^{\pi}(t) \left(Q_j^{\pi}(t) - Q_i^{\pi}(t) + \nu \hat{c}_{ij}(t)\right)\right]$$

where, $B := \frac{1}{2} \sum_{i \in \mathcal{N}} [(\sum_{j \in \mathcal{N}_i} \mu_{ij}^{max})^2 + (\sum_{j : i \in \mathcal{N}_j} \mu_{ji}^{max})^2] + \frac{1}{2} \mathbb{E}[a(t)^2]$ is a constant that depends only on edge capacities and the second moment of arrivals.

Now, we derive the Drift Plus Optimistic Penalty (DPOP) policy π_D by minimizing this bound as shown below. Define the set of all feasible transmissions as $\mathcal{M} := \{\boldsymbol{\mu} : \forall (i,j) \in \mathcal{E},\ 0 \leq \mu_{ij} \leq \mu_{ij}^{max}\}$. Given the backlogs $\{Q_i^{\pi_D}(t)\}_{i \in \mathcal{N}}$ and the lower-confidence-bound cost estimates $\{\hat{c}_{ij}(t)\}_{(i,j) \in \mathcal{E}}$ at time t, the policy π_D picks edge transmission values $\boldsymbol{\mu}^{\pi_D}(t) := \{\mu_{ij}^{\pi_D}(t)\}_{(i,j) \in \mathcal{E}}$ by greedily minimizing the bound on $\hat{L}^{\pi}(t)$. Hence, ignoring the uncontrollable terms, we get

$$\boldsymbol{\mu}^{\pi_D}(t) = \max_{\boldsymbol{\mu} \in \mathcal{M}} \sum_{(i,j) \in \mathcal{E}} \mu_{ij} \left(Q_i^{\pi_D}(t) - Q_j^{\pi_D}(t) - \nu \hat{c}_{ij}(t)\right). \tag{7}$$

Algorithm 1 Drift Plus Optimistic Penalty (DPOP) Algorithm.

Input: Network G, Parameters β, δ, and ν.

at $t = 0$ **do**

1: Send a null packet on each edge, $\mu_{ij}^{\pi_D}(0) = 1,\ \forall (i,j)$.

2: Observe noisy costs $\widetilde{c}_{ij}(0),\ \forall (i,j) \in \mathcal{E}$.

3: Set $\bar{c}_{ij}(1) = \widetilde{c}_{ij}(0),\ \forall (i,j) \in \mathcal{E}$.

4: Set $N_{ij}(1) = 1,\ \forall (i,j) \in \mathcal{E}$.

for $t = 1, 2, ..., T$ **do**

5: Calculate cost estimates $\{\hat{c}_{ij}(t)\}_{(i,j) \in \mathcal{E}}$ using (6).

6: Pick transmissions $\boldsymbol{\mu}^{\pi_D}(t)$ according to (7).

7: Send null packets on each edge $(i,j) \in \mathcal{E}$ where $\mu_{ij}^{\pi_D}(t) > 0$ but queue $Q_i^{\pi_D}(t)$ does not have sufficient packets to support the planned transmission $\mu_{ij}^{\pi_D}(t)$.

8: Update queue lengths according to (1), (2), and (3).

9: Observe noisy costs $\widetilde{c}_{ij}(t)$ for all $(i,j) : \mu_{ij}^{\pi_D}(t) > 0$.

10: $\forall (i,j) : \mu_{ij}^{\pi_D}(t) > 0$, update average costs $\bar{c}_{ij}(t+1)$

$$\bar{c}_{ij}(t+1) = \bar{c}_{ij}(t) + \frac{\widetilde{c}_{ij}(t) - \bar{c}_{ij}(t)}{N_{ij}(t) + 1}.$$

11: $\forall (i,j) : \mu_{ij}^{\pi_D}(t) > 0,\ N_{ij}(t+1) = N_{ij}(t) + 1$.

end for

We present the DPOP policy in Algorithm 1. We start with an initial exploration phase at $t = 0$ before any packets arrive. In this phase, we send a null packet on each edge (line 1) to receive an initial cost observation (line 2). Using this observation, we initialize the average cost estimates $\bar{c}_{ij}(t)$ (line 3) and $N_{ij}(t)$ (line 4) for $t = 1$. New packets arrive starting from $t = 1$. At each time $t \in \{1, ..., T\}$, we calculate the optimistic estimates $\hat{c}_{ij}(t)$ (line 5) and make transmission decisions $\mu_{ij}^{\pi_D}(t)$ by greedily minimizing the bound on Lyapunov drift-plus-optimistic-penalty (line 6). It is possible that for some edges, their neighboring nodes do not have sufficient packets to support the planned transmissions. On those edges, we send null packets such that the total transmission is no more than the planned transmission (line 7). This allows us to get observations for all the chosen edges. Note that the null packets do not contribute to the queue evolution but they do incur transmission costs just like normal packets. We then perform the planned transmissions, which updates the queue backlogs (line 8). We receive noisy cost observations for each edge on which we transmitted packets (line 9). Finally, we update the cost averages $\bar{c}_{ij}(t)$ (line 10) and the number of observations $N_{ij}(t)$ (line 11).

V. Regret Analysis

In this section, we analyze the regret of DPOP policy π_D and derive an upper-bound for the regret. We formally present the policy's regret performance in the following theorem.

Theorem 2: (DPOP Regret) The DPOP policy π_D has regret

$$R^{\pi_D}(T) = O(T^{2/3} + \sqrt{T}\log T)$$

with parameters $\beta > 4\sigma^2$, $\delta = T^{-\frac{2\sigma^2}{\beta - 2\sigma^2}}$, and $\nu = T^{1/3}$.

Proof of Theorem 2: The proof outline is as follows. We first decompose π_D's regret into four components in Lemma 1. We then bound each of these components individually in Propositions 1 to 4.

We first define event A as the event that the true edge costs c_{ij}'s are within the confidence interval of our estimates $\bar{c}_{ij}(t)$'s at all slots t and at all edges.

$$A := \left\{ \forall t, \forall (i,j),\ |c_{ij} - \bar{c}_{ij}(t)| \leq \sqrt{\frac{\beta \log(t/\delta)}{N_{ij}(t)}} \right\}.$$

Denote the compliment of event A by $\bar{A}$. Now, in Lemma 1, we decompose the regret into four components $R_1^{\pi_D}(T), ..., R_4^{\pi_D}(T)$ defined as follows.

$$R_1^{\pi_D}(T) := \sum_{t=1}^{T} \sum_{(i,j)\in\mathcal{E}} \mathbb{E}\left[\mu_{ij}^{\pi_D}(t)(c_{ij} - \hat{c}_{ij}(t)) \mid A\right],$$

$$R_2^{\pi_D}(T) := T \sum_{(i,j)\in\mathcal{E}} \mu_{ij}^{max} C_{max}\, \mathbb{P}[\bar{A}],$$

$$R_3^{\pi_D}(T) := \sum_{t=1}^{T} \sum_{(i,j)\in\mathcal{E}} \mathbb{E}[(\mu_{ij}^{\pi_D}(t) - \mu_{ij}^{stat})\hat{c}_{ij}(t) \mid A],$$

$$R_4^{\pi_D}(T) := C_B \sum_{i\in\mathcal{N}} \mathbb{E}[Q_i^{\pi_D}(T)].$$

The first component $R_1^{\pi_D}(T)$ corresponds to the gap between the true and estimated costs. The second component $R_2^{\pi_D}(T)$ captures the probability that the true cost is outside the confidence interval of our estimate. The third component $R_3^{\pi_D}(T)$ corresponds to the gap between policy's transmission cost and the static cost. Finally, the component $R_4^{\pi_D}(T)$ corresponds to the backlog cost.

Lemma 1: (Regret Decomposition) The regret $R^{\pi_D}(T)$ of the DPOP policy π_D can be decomposed as

$$R^{\pi_D}(T) \leq R_1^{\pi_D}(T) + R_2^{\pi_D}(T) + R_3^{\pi_D}(T) + R_4^{\pi_D}(T).$$

We prove Lemma 1 in Appendix B. The proof involves using the regret upper bound from Corollary 1 and conditioning the regret on events A and $\bar{A}$. We ignore the cost of exploration phase at $t = 0$ as this only adds a constant term to the regret. Next, we bound these four components in the following propositions.

- *Proposition 1:* $R_1^{\pi_D}(T) = O(\sqrt{T}\log T)$.
- *Proposition 2:* $R_2^{\pi_D}(T) = O(1)$ with $\delta = T^{\frac{-2\sigma^2}{\beta - 2\sigma^2}}$, $\beta > 4\sigma^2$.
- *Proposition 3:* $R_3^{\pi_D}(T) = O(T^{2/3})$ with $\nu = T^{1/3}$.
- *Proposition 4:* $R_4^{\pi_D}(T) = O(T^{2/3})$ with $\nu = T^{1/3}$.

We prove Propositions 1 to 4 in Appendices C to F respectively. Combining Lemma 1 and Propositions 1 to 4, we can see that the overall regret of the policy is of order $O(T^{2/3} + \sqrt{T}\log T)$. This proves Theorem 2 and concludes the regret analysis. □

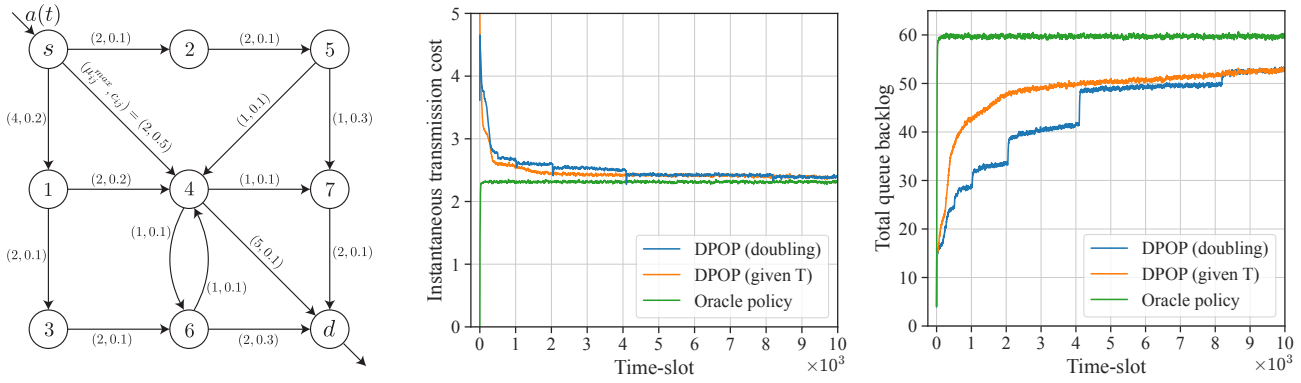
VI. Simulation Results

We evaluate[3] our policy on a queueing network with 9 nodes and 15 edges shown in Fig. 1a. The edge capacities and transmission costs are shown as tuples (μ_{ij}^{max}, c_{ij}) marked on their respective edges in the figure. For this network, we can calculate the stability region using (4) to be $\Lambda(G) = [0, 8]$. New packets arrive at s according to a Poisson process with mean $\lambda \in \Lambda(G)$. The cost observations are corrupted by independent random variables uniformly distributed in $[-\sigma, \sigma]$.

A. Queue Backlog and Transmission Cost Performance

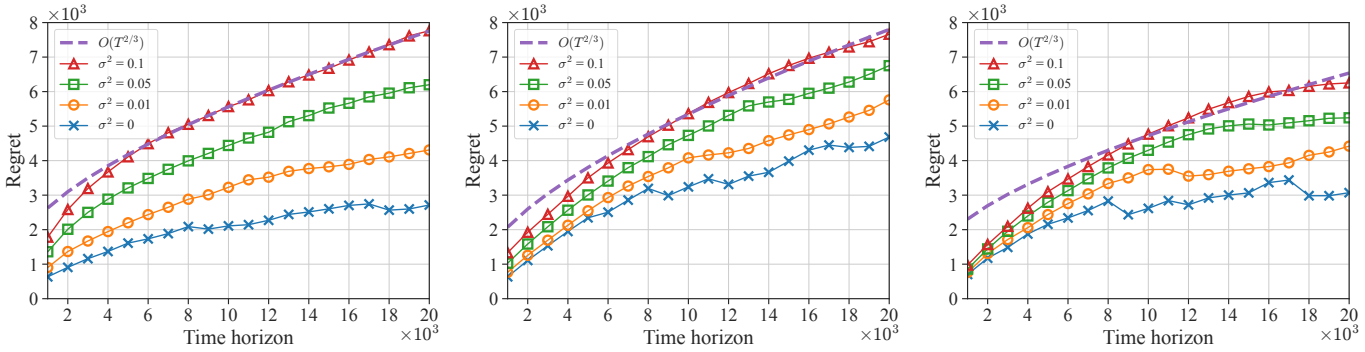
We simulate the DPOP policy for $T = 10000$, $\lambda = 4$, and $\sigma^2 = 0.05$. We use Theorem 2 to choose tuning parameters $\beta = 4.5\sigma^2$, $\delta = T^{-2\sigma^2/(\beta - 2\sigma^2)}$, and $\nu = T^{1/3}$. This choice of tuning parameters requires the knowledge of T, which may not be available in some practical scenarios. Hence, we also simulate the policy for the case of unknown T using the doubling trick [29]. Here, we pick the tuning parameters according to an estimated time horizon $\hat{T}$ (initialized as $\hat{T} = 2$), which we double whenever the current time-slot t crosses the current estimate, i.e. $\hat{T} \leftarrow 2\hat{T}$ if $t > \hat{T}$. Finally, to benchmark our policy, we also simulate an oracle policy that has access to the true transmission costs. The oracle policy is similar to the DPOP policy described in Algorithm 1 except that in line 6, it uses the true costs c_{ij}'s instead of the optimistic estimates $\hat{c}_{ij}(t)$'s.

[3]See implementations at https://github.com/sathwikchadaga/optimistic-dpp.



(a) Network topology showing (μ_{ij}^{max}, c_{ij}). (b) Transmission cost $\sum_{(i,j)\in\mathcal{E}} \mu_{ij}^{\pi}(t)c_{ij}$. (c) Total queue backlog $\sum_{i\in\mathcal{N}} Q_i^{\pi}(t)$.

Fig. 1: Network topology (left) and the corresponding total transmission costs (middle) and backlogs (right) plotted against t.



(a) Regret for $\lambda = 2$. (b) Regret for $\lambda = 4$. (c) Regret for $\lambda = 6$.

Fig. 2: Upper bound of regret $R^{\pi}(T)$ plotted as a function of time horizon T for various arrival rates λ and noise levels σ.

Fig. 1b shows the plot of resulting total transmission costs $\sum_{(i,j)\in\mathcal{E}} \mu_{ij}^{\pi}(t)c_{ij}$ as a function of time-slots t. The figure has three plots corresponding to the DPOP policy with doubling trick (unknown T), DPOP policy with known T, and the oracle policy. From this figure, we can see that the DPOP policy's cost is large initially and reduces later to eventually approach the oracle policy. This shows the policy's learning process and its ability to explore and learn the low cost routes. Similarly, Fig. 1c shows the plot of total queue backlogs $\sum_{i\in\mathcal{N}} Q_i^{\pi}(t)$ as a function of time-slots t. From this figure, we can see that the queue backlogs do not grow indefinitely, which demonstrates the DPOP policy's ability to stabilize the network.

Further, compared to the oracle policy, we can see that the DPOP policy has lower queue backlogs but has higher transmission costs. This is due to the exploratory bias in DPOP's decision making. Indeed, as part of the exploration, it sometimes uses higher cost paths and hence reduces some queue backlogs. Finally, we can also see that the performance of DPOP policy with doubling (for unknown T) is close to the performance of DPOP policy with known T. This shows that the policy can also be used when the value of T is not known in advance.

B. Regret Comparison for Varying Noise and Arrival Rates

We evaluate the DPOP policy's regret for different values of λ and σ. To calculate the regret, we use the upper bound from Corollary 1. We obtain the policy costs $C^{\pi}(T)$ from simulations and calculate the solution to the static optimization problem $\mathcal{P}$ using a solver. Employing the upper bound, we calculate the regret $R^{\pi_D}(T)$ as the gap between policy cost $C^{\pi}(T)$ and T times the static optimal cost.

Fig. 2 shows the resulting regret plotted as a function of $T = \{1000, 2000, ..., 20000\}$. Specifically, figures 2a, 2b, and 2c show the regrets for $\lambda = 2, 4$, and 6 respectively. In each of these figures, we show the DPOP policy's regrets for different values of noise parameter $\sigma^2 = \{0.01, 0.05, 0.1\}$. From the plots, we can see that the regret increases as the noise level increases, as expected. Further, we can see that the regret for $\lambda = 6$ is slightly lower than the regrets for $\lambda = 2, 4$. This is because the rate $\lambda = 6$ is close to the network's max-flow of 8. Due to this heavy loading, all the policies, including the static optimal policy, are forced to use almost all of the existing paths. Hence, the regret compared to the static optimal policy is small. Further, we also plot a $O(T^{2/3})$ curve fitted to the DPOP regret of $\sigma^2 = 0.1$ (we omit other fits to avoid crowding). This demonstrates that the DPOP policy indeed has a sub-linear regret. Finally, we clarify that the oracle policy has a non-zero regret because, even though it knows the true costs, it still does not know the arrival rate. Therefore, due to the stochastic nature of arrivals, the gap between oracle policy's cost and the static cost is non-zero.

VII. Conclusion

We considered the problem of making scheduling and routing decisions in queueing networks, where the transmission costs are initially unknown. To design policies that optimize both throughput and cost simultaneously, we defined a novel cost metric in terms of transmission costs and queue backlogs. We showed that any policy's cost is lower bounded by the solution to a static optimization problem. Using this lower-bound, we defined any control policy's regret as the difference between policy's total cost and T times the solution to static optimization problem. Further, using techniques from Lyapunov theory and multi-arm bandits, we designed the Drift Plus Optimistic Penalty Policy. We showed that this policy achieves a sub-linear regret of order $O(T^{2/3} + \sqrt{T}\log T)$ and evaluated its performance using simulations. Finally, the transmission costs in our paper are independent of the queue backlog states. Designing control policies when costs are a function of queue states is a potential future research direction.

Acknowledgment

This work was supported by NSF grants CNS-2148183 and CNS-2106268. We also thank Xinzhe Fu for the helpful discussions.

Appendix A
Proof of Theorem 1 (Static Lower Bound)

Theorem 1: $\inf_{\pi^* \in \Pi^*} \mathbb{E}[C^{\pi^*}(T)] \geq T \sum_{(i,j)\in\mathcal{E}} \mu_{ij}^{stat} c_{ij}$ for $C_B \geq C_L$ where, recall $\{\mu_{ij}^{stat}\}_{(i,j)\in\mathcal{E}}$ is the solution to $\mathcal{P}$.

Proof: We first define some variables that will be useful for the proof. Recall that the number of packets transmitted on any edge (i,j) under policy π^* at time t is $\tilde{\mu}_{ij}^{\pi^*}(t)$. Among these packets, let $\hat{\mu}_{ij}^{\pi^*}(t)$ be the number of packets that got delivered to the destination by the end of time horizon T. We define the average effective rate as $\bar{\mu}_{ij}^{\pi^*} = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \hat{\mu}_{ij}^{\pi^*}(t)\right]$ and define the average final backlog as $\bar{Q}_T^{\pi^*} = \mathbb{E}\left[\sum_{i\in\mathcal{N}} Q_i^{\pi^*}(T)\right]$. We show certain properties of these quantities in the lemma below.

Lemma 2: The average effective rates $\{\bar{\mu}_{ij}^{\pi^*}\}_{(i,j)\in\mathcal{E}}$ and the average final backlog $\bar{Q}_T^{\pi^*}$ satisfy

$$\lambda = \sum_{j\in\mathcal{N}_s} \bar{\mu}_{sj}^{\pi^*} + \bar{Q}_T^{\pi^*}/T, \tag{8}$$

$$\sum_{j:i\in\mathcal{N}_j} \bar{\mu}_{ji}^{\pi^*} = \sum_{j\in\mathcal{N}_i} \bar{\mu}_{ij}^{\pi^*}, \ \forall i \in \mathcal{N}\setminus\{s,d\}. \tag{9}$$

Lemma 2 can be proved by exploiting the fact that $\bar{\mu}_{ij}^{\pi^*}$ satisfies flow constraints. We omit the proof for brevity. Now, we express the best policy's cost in terms of $\{\bar{\mu}_{ij}^{\pi^*}\}_{(i,j)\in\mathcal{E}}$ and $\bar{Q}_T^{\pi^*}$. Since the packets counted by $\hat{\mu}_{ij}^{\pi^*}(t)$ is a subset of the packets counted by $\tilde{\mu}_{ij}^{\pi^*}(t)$, we have $\tilde{\mu}_{ij}^{\pi^*}(t) \geq \hat{\mu}_{ij}^{\pi^*}(t)$. Hence, using $C_B \geq C_L$, we have

$$\begin{aligned}\mathbb{E}[C^{\pi^*}(T)] &\geq \mathbb{E}\sum_{t=1}^{T}\sum_{(i,j)\in\mathcal{E}} \hat{\mu}_{ij}^{\pi^*}(t)c_{ij} + C_L\,\mathbb{E}\sum_{i\in\mathcal{N}} Q_i^{\pi^*}(T)\\ &= T\sum_{(i,j)\in\mathcal{E}} \bar{\mu}_{ij}^{\pi^*} c_{ij} + C_L \bar{Q}_T^{\pi^*}\end{aligned} \tag{10}$$

Now, if the average final backlog $\bar{Q}_T^{\pi^*}$ is zero, then the average effective rates $\{\bar{\mu}_{ij}^{\pi^*}\}_{(i,j)\in\mathcal{E}}$ lie within the feasibility region of $\mathcal{P}$ and the result holds trivially. However, when there are undelivered packets at the end of time horizon, we can see from (8) that the source node rates $\{\bar{\mu}_{sj}^{\pi^*}\}_{j\in\mathcal{N}_s}$ violate the first constraint of $\mathcal{P}$. Therefore, this requires a careful analysis.

When there are undelivered packets left at T, the policy incurs a non-zero final backlog cost of $C_L\bar{Q}_T^{\pi^*}$. To further bound this backlog cost, we show in Lemma 3 that there exists a feasible flow $\{f_{ij}\}_{(i,j)\in\mathcal{E}}$ that can be used to route these backlogged packets to the destination only using the residual edge capacities. Since this flow uses only the residual capacities, we will be able to show that the sum of flows $\{\bar{\mu}_{ij}^{\pi^*} + f_{ij}\}_{(i,j)\in\mathcal{E}}$ lies within the feasibility region of $\mathcal{P}$. This will allow us to bound the cost further and conclude the proof.

Lemma 3: There exists a cycle-free flow $\{f_{ij},\ \forall (i,j)\in\mathcal{E}\}$ that satisfies the following constraints.

$$\sum_{j\in\mathcal{N}_s} f_{sj} = \bar{Q}_T^{\pi^*}/T, \tag{11}$$

$$\sum_{j:i\in\mathcal{N}_j} f_{ji} = \sum_{j\in\mathcal{N}_i} f_{ij},\ \forall i \in \mathcal{N}\setminus\{s,d\}, \tag{12}$$

$$0 \leq f_{ij} \leq \mu_{ij}^{max} - \bar{\mu}_{ij}^{\pi^*},\ \forall (i,j)\in\mathcal{E}. \tag{13}$$

Note that (13) ensures that the flow $\{f_{ij}\}_{(i,j)\in\mathcal{E}}$ can be routed on the residual link capacities after accounting for the traffic $\{\bar{\mu}_{ij}^{\pi^*}\}_{(i,j)\in\mathcal{E}}$ that actually reached the destination.

Proof of Lemma 3: Define the residual network G' as follows. G' is the same as G in all properties except that its edge capacities are reduced as $\mu'^{max}_{ij} = \mu_{ij}^{max} - \bar{\mu}_{ij}^{\pi^*}$. Now, to show that there exists a flow that satisfies (11), (12), and (13), it is enough to show that $\bar{Q}_T^{\pi^*}/T \leq \text{max-flow}(G')$. Indeed,

$$\begin{aligned}\text{max-flow}(G') &= \text{min-cut-capacity}(G')\\ &= \text{min-cut-capacity}(G) - \sum_{j\in\mathcal{N}_s}\bar{\mu}_{sj}^{\pi^*}\\ &= \text{max-flow}(G) - \sum_{j\in\mathcal{N}_s}\bar{\mu}_{sj}^{\pi^*}\\ &\geq \lambda - \sum_{j\in\mathcal{N}_s}\bar{\mu}_{sj}^{\pi^*} = \bar{Q}_T^{\pi^*}/T.\end{aligned}$$

where, the first equality is due to the max-flow min-cut theorem, the second equality is due to the fact that any cut in the graph will have equal total flow, the inequality is due to the fact that $\lambda \in \Lambda(G)$, and the final equality is from Lemma 2's equation (8). Therefore, we have shown that there exists a valid flow such that (11), (12), and (13) are satisfied. Let such a flow be $\{f_{ij}\}_{(i,j)\in\mathcal{E}}$. Further, from the flow decomposition theorem, we know that any flow can be decomposed into s-d flows and cycles. This means that, by decomposing and removing any cycles, the flow $\{f_{ij}\}_{(i,j)\in\mathcal{E}}$ can easily be made cycle-free while keeping all the previous properties. □

Proof of Theorem 1, Continued: In Lemma 3, we showed the existence of a cycle-free flow $\{f_{ij}\}_{(i,j)\in\mathcal{E}}$. Now, recall that C_L was defined as the per-packet-cost of the longest s-d path in the network. Since $\{f_{ij}\}_{(i,j)\in\mathcal{E}}$ is a cycle-free flow

that adheres to flow conservation at every node (12), each of the packets in this flow can at most incur a total cost of C_L. Further, from (11), we know that the flow volume or the amount of packets in the flow is $\bar{Q}_T^{\pi^*}/T$. Therefore, we have $\sum_{(i,j)\in\mathcal{E}} f_{ij}c_{ij} \leq C_L\bar{Q}_T^{\pi^*}/T$. Plugging this in (10), we can bound the best policy's cost as $\mathbb{E}[C^{\pi^*}(T)] \geq T\sum_{(i,j)\in\mathcal{E}}(\bar{\mu}_{ij}^{\pi^*}+f_{ij})c_{ij}$. Now, to bound the cost further, we can easily verify from Lemmas 2 and 3 that $\{\bar{\mu}_{ij}^{\pi^*}+f_{ij}\}_{(i,j)\in\mathcal{E}}$ lies within the feasibility region of $\mathcal{P}$. Therefore, we get

$$\mathbb{E}[C^{\pi^*}(T)] \geq T\sum_{(i,j)\in\mathcal{E}}(\bar{\mu}_{ij}^{\pi^*}+f_{ij})c_{ij} \geq T\sum_{(i,j)\in\mathcal{E}}\mu_{ij}^{stat}c_{ij}. \quad \square$$

Appendix B
Proof of Lemma 1 (Regret Decomposition)

From Corollary 1, we can bound the regret as $R^{\pi_D}(T) \leq \sum_t\sum_{ij\in\mathcal{E}}\mathbb{E}\left[(\mu_{ij}^{\pi_D}(t)-\mu_{ij}^{stat})c_{ij}\right] + C_B\sum_i\mathbb{E}[Q_i^{\pi_D}(T)]$. Now, we analyze the term inside first summation by conditioning it on the event A and its compliment $\bar{A}$.

$$\mathbb{E}[(\mu_{ij}^{\pi_D}(t)-\mu_{ij}^{stat})c_{ij}] = \mathbb{E}\left[(\mu_{ij}^{\pi_D}(t)-\mu_{ij}^{stat})c_{ij} \mid A\right]\mathbb{P}[A] + \mathbb{E}\left[(\mu_{ij}^{\pi_D}(t)-\mu_{ij}^{stat})c_{ij} \mid \bar{A}\right]\mathbb{P}[\bar{A}]. \quad (14)$$

We first bound the expectation conditioned on A as follows. Under event A, by definition, $\forall t$, $\forall(i,j)$, $c_{ij} \geq \bar{c}_{ij}(t) - \sqrt{\beta\log(t/\delta)/N_{ij}(t)} = \hat{c}_{ij}(t)$. Also since $\mu_{ij}^{stat} \geq 0$, we get $(\mu_{ij}^{\pi_D}(t)-\mu_{ij}^{stat})(c_{ij}-\hat{c}_{ij}(t)) \leq \mu_{ij}^{\pi_D}(t)(c_{ij}-\hat{c}_{ij}(t))$. Hence, Adding and subtracting the optimistic estimates $\hat{c}_{ij}(t)$'s to c_{ij}'s inside $\mathbb{E}[(\mu_{ij}^{\pi_D}(t)-\mu_{ij}^{stat})c_{ij} \mid A]$, we get

$$\mathbb{E}[(\mu_{ij}^{\pi_D}(t)-\mu_{ij}^{stat})c_{ij} \mid A] \leq \mathbb{E}\left[\mu_{ij}^{\pi_D}(t)(c_{ij}-\hat{c}_{ij}(t)) \mid A\right] + \mathbb{E}\left[(\mu_{ij}^{\pi_D}(t)-\mu_{ij}^{stat})\hat{c}_{ij}(t) \mid A\right].$$

Now, we bound the expectation conditioned on $\bar{A}$. Since $\mu_{ij}^{\pi_D}(t)-\mu_{ij}^{stat} \leq \mu_{ij}^{\pi_D}(t) \leq \mu_{ij}^{max}$ and since $c_{ij}\in[0,C_{max}]$,

$$\mathbb{E}\left[(\mu_{ij}^{\pi_D}(t)-\mu_{ij}^{stat})c_{ij} \mid \bar{A}\right] \leq C_{max}\mu_{ij}^{max}.$$

Plugging these back in (14) and using $\mathbb{P}[A]\leq 1$,

$$\mathbb{E}\left[(\mu_{ij}^{\pi_D}(t)-\mu_{ij}^{stat})c_{ij}\right] \leq \mathbb{E}\left[\mu_{ij}^{\pi_D}(t)(c_{ij}-\hat{c}_{ij}(t)) \mid A\right] + \mathbb{E}\left[(\mu_{ij}^{\pi_D}(t)-\mu_{ij}^{stat})\hat{c}_{ij}(t) \mid A\right] + C_{max}\mu_{ij}^{max}\,\mathbb{P}[\bar{A}].$$

Plugging this in the $R^{\pi_D}(T)$ bound completes the proof. $\square$

Appendix C
Proof of Proposition 1

Proposition 1: $R_1^{\pi_D}(T) = O(\sqrt{T}\log T)$.

Proof: Starting from the definition of $R_1^{\pi_D}(T) = \sum_t\sum_{(i,j)\in\mathcal{E}}\mathbb{E}\left[\mu_{ij}^{\pi_D}(t)(c_{ij}-\hat{c}_{ij}(t)) \mid A\right]$, adding and subtracting $\bar{c}_{ij}(t)$'s to c_{ij}'s, we obtain

$$R_1^{\pi_D}(T) \leq \sum_{t=1}^{T}\sum_{(i,j)\in\mathcal{E}}\mathbb{E}[\mu_{ij}^{\pi_D}(t)(|c_{ij}-\bar{c}_{ij}(t)| + |\bar{c}_{ij}(t)-\hat{c}_{ij}(t)|) \mid A].$$

Under the event A, we have $|c_{ij}-\bar{c}_{ij}(t)| \leq \sqrt{\beta\log(t/\delta)/N_{ij}(t)} \leq \sqrt{\beta\log(T/\delta)/N_{ij}(t)}$. Also, from the definition of $\hat{c}_{ij}(t)$, we have $|\hat{c}_{ij}(t)-\bar{c}_{ij}(t)| = \sqrt{\beta\log(t/\delta)/N_{ij}(t)} \leq \sqrt{\beta\log(T/\delta)/N_{ij}(t)}$. Hence,

$$R_1^{\pi_D}(T) \leq 2\sum_{t=1}^{T}\sum_{(i,j)\in\mathcal{E}}\mathbb{E}\left[\mu_{ij}^{\pi_D}(t)\sqrt{\frac{\beta\log(T/\delta)}{N_{ij}(t)}} \;\middle|\; A\right]. \quad (15)$$

Lemma 4: $\sum_{t=1}^{T}\sum_{ij\in\mathcal{E}}\mu_{ij}^{\pi_D}(t)/\sqrt{N_{ij}(t)} = O(\sqrt{T\log T})$.

Plugging Lemma 4's result in (15), we can further bound $R_1^{\pi_D}(T)$ and obtain the desired result. To conclude the proposition's proof, we are only left with proving Lemma 4.

Proof of Lemma 4: Let $x_{ij}(t) := \mu_{ij}^{\pi_D}(t)/\mu_{ij}^{max}$. Expressing in terms of $x_{ij}(t)$ and using Cauchy–Schwarz,

$$\sum_{t=1}^{T}\sum_{(i,j)\in\mathcal{E}}\frac{\mu_{ij}^{\pi_D}(t)}{\sqrt{N_{ij}(t)}} = \sum_{t=1}^{T}\sum_{(i,j)\in\mathcal{E}}\frac{\mu_{ij}^{max}x_{ij}(t)}{\sqrt{N_{ij}(t)}} \leq \sqrt{\sum_{t=1}^{T}\sum_{(i,j)\in\mathcal{E}}(\mu_{ij}^{max})^2}\sqrt{\sum_{t=1}^{T}\sum_{(i,j)\in\mathcal{E}}\frac{x_{ij}^2(t)}{N_{ij}(t)}}.$$

Further, by definition, $x_{ij}(t)\in[0,1]$ and $N_{ij}(t)\geq 1$, hence $\frac{x_{ij}^2(t)}{N_{ij}(t)}\in[0,1]$. Using the fact $\forall y\in[0,1]$, $y\leq 2\log(1+y)$, we can see that $\frac{x_{ij}^2(t)}{N_{ij}(t)} \leq 2\log\left(1+\frac{x_{ij}^2(t)}{N_{ij}(t)}\right)$. Plugging this in the above inequality,

$$\sum_{t=1}^{T}\sum_{(i,j)\in\mathcal{E}}\frac{\mu_{ij}^{\pi_D}(t)}{\sqrt{N_{ij}(t)}} \leq B_1\sqrt{T\sum_{(i,j)\in\mathcal{E}}\sum_{t=1}^{T}\log\left(1+\frac{x_{ij}^2(t)}{N_{ij}(t)}\right)} \quad (16)$$

where, $B_1 = \sqrt{2\sum_{(i,j)\in\mathcal{E}}(\mu_{ij}^{max})^2}$. Now, using $N_{ij}(t)$'s update equation, $N_{ij}(t+1) = N_{ij}(t) + \mathbb{I}[\mu_{ij}^{\pi}(t)>0] \geq N_{ij}(t) + x_{ij}^2(t) = N_{ij}(t)(1+\frac{x_{ij}^2(t)}{N_{ij}(t)})$. Taking $\log(\cdot)$ and telescoping over time $t=1,2,...,T-1$, we have $\forall(i,j)$,

$$\log N_{ij}(T+1) \geq \log N_{ij}(1) + \sum_{t=1}^{T}\log\left(1+\frac{x_{ij}^2(t)}{N_{ij}(t)}\right).$$

Finally, using the fact that $N_{ij}(1)=1$ and $N_{ij}(T+1)\leq T+1$ for all (i,j), we have $\sum_{t=1}^{T}\log\left(1+\frac{x_{ij}^2(t)}{N_{ij}(t)}\right) \leq \log(T+1)$. Plugging this back in (16) completes the proof. $\square$

Appendix D
Proof of Proposition 2

Proposition 2: $R_2^{\pi_D}(T) = O(1)$ with $\delta = T^{\frac{-2\sigma^2}{\beta-2\sigma^2}}$, $\beta > 4\sigma^2$.

Proof: To bound $R_2^{\pi_D}(T)$, we analyze the probability term $\mathbb{P}(\bar{A})$. Using the union bound,

$$\mathbb{P}(\bar{A}) = \mathbb{P}\left[\exists t,\exists(i,j) : |c_{ij}-\bar{c}_{ij}(t)| > \sqrt{\frac{\beta\log(t/\delta)}{N_{ij}(t)}}\right] \leq \sum_{t=1}^{T}\sum_{(i,j)\in\mathcal{E}}\mathbb{P}\left[|c_{ij}-\bar{c}_{ij}(t)| > \sqrt{\frac{\beta\log(t/\delta)}{N_{ij}(t)}}\right]. \quad (17)$$

Define $\mathcal{T}_{ij}(t) = \{\tau\in\{0,1,...,t-1\} : \mu_{ij}^{\pi}(\tau)>0\}$ as the set of time-slots until t in which a feedback value was received. It follows that $|\mathcal{T}_{ij}(t)| = N_{ij}(t)$. Moreover, the noisy

feedback values received until time t for each edge (i,j) are $\forall \tau \in \mathcal{T}_{ij}(t)$, $\widetilde{c}_{ij}(\tau) = c_{ij} + \eta_{ij}(\tau)$. Hence, for $t \geq 1$, the average cost estimate $\bar{c}_{ij}(t)$ is given by

$$\bar{c}_{ij}(t) = \frac{\sum_{\tau \in \mathcal{T}_{ij}(t)} \widetilde{c}_{ij}(\tau)}{|\mathcal{T}_{ij}(t)|} = c_{ij} + \frac{\sum_{\tau \in \mathcal{T}_{ij}(t)} \eta_{ij}(\tau)}{N_{ij}(t)}.$$

where, the second equality is valid since $N_{ij}(t) \geq 1$, $\forall (i,j)$ and $t \geq 1$. Denoting $\theta_t := \sqrt{\beta \log(t/\delta)/N_{ij}(t)}$, we now have

$$\mathbb{P}\Big[|c_{ij} - \bar{c}_{ij}(t)| > \theta_t\Big] = \mathbb{P}\Big[\frac{|\sum_{\tau \in \mathcal{T}_{ij}(t)} \eta_{ij}(\tau)|}{N_{ij}(t)} > \sqrt{\frac{\beta \log \frac{t}{\delta}}{N_{ij}(t)}}\Big] \leq \sum_{n=1}^{t-1} \mathbb{P}\Big[|\sum_{\tau \in \mathcal{T}_{ij}(t)} \eta_{ij}(\tau)| > \sqrt{n\beta \log(t/\delta)}\Big]$$

where, in the last inequality we have conditioned on number of observations $N_{ij}(t) = |\mathcal{T}_{ij}(t)| = n$. Now, since $\eta_{ij}(t)$ are independent σ-sub-Gaussian, Hoeffding's bound [28] gives us

$$\mathbb{P}\Big[|c_{ij} - \bar{c}_{ij}(t)| > \theta_t\Big] \leq \sum_{n=1}^{t-1} 2e^{-\frac{\beta}{2\sigma^2} \log t/\delta} = 2(t/\delta)^{1-\beta/2\sigma^2}.$$

Plugging this in (17), taking $\beta > 4\sigma^2$, and $\delta = T^{-\frac{2\sigma^2}{\beta - 2\sigma^2}}$, we have $\mathbb{P}[\bar{A}] \leq \sum_t \sum_{ij \in \mathcal{E}} 2(t/\delta)^{1-\beta/2\sigma^2} = O(1/T)$. Hence, $R_2^{\pi_D}(T) = T \sum_{(i,j) \in \mathcal{E}} \mu_{ij}^{max} C_{max}\, \mathbb{P}[\bar{A}] = O(1)$. □

Appendix E
Proof of Proposition 3

Proposition 3: $R_3^{\pi_D}(T) = O(T^{2/3})$, with $\nu = T^{1/3}$.

Proof: We start with the following lemma to bound the conditional drift-plus-optimistic-penalty.

Lemma 5: Given $\{Q_i^{\pi_D}(t)\}_{i \in \mathcal{N}}$ and $\{\hat{c}_{ij}(t)\}_{(i,j) \in \mathcal{E}}$, policy π_D's conditional drift-plus-optimistic-penalty is bounded as

$$\mathbb{E}_{|Q,\hat{c}}\Big[\Delta\Phi^{\pi_D}(t) + \nu \sum_{(i,j) \in \mathcal{E}} \mu_{ij}^{\pi_D}(t) \hat{c}_{ij}(t)\Big] \leq B + \nu \sum_{(i,j) \in \mathcal{E}} \mu_{ij}^{stat} \hat{c}_{ij}(t).$$

where, B is a constant independent of T and we denote $\mathbb{E}_{|Q,\hat{c}}[\cdot] := \mathbb{E}[\cdot \mid \{Q_i^{\pi_D}(t)\}_{i \in \mathcal{N}}, \{\hat{c}_{ij}(t)\}_{(i,j) \in \mathcal{E}}]$.

Taking $\mathbb{E}[\cdot|A]$ on both sides of Lemma 5's result and using total law of expectations, we get $\mathbb{E}[\Delta\Phi^{\pi_D}(t) + \nu \sum_{ij \in \mathcal{E}} \mu_{ij}^{\pi_D}(t) \hat{c}_{ij}(t) \mid A] \leq B + \nu \sum_{ij \in \mathcal{E}} \mathbb{E}[\mu_{ij}^{stat} \hat{c}_{ij}(t) \mid A]$. Rearranging and summing over $t = 1, 2, ..., T$, we get

$$R_3^{\pi_D}(T) = \sum_{t=1}^{T} \sum_{(i,j) \in \mathcal{E}} \mathbb{E}[(\mu_{ij}^{\pi_D}(t) - \mu_{ij}^{stat}) \hat{c}_{ij}(t) \mid A] \leq BT/\nu - \mathbb{E}[\Phi^{\pi_D}(T+1) \mid A]/\nu \leq BT/\nu.$$

where, the last inequality is because $\Phi^{\pi_D}(t) \geq 0, \forall t$. Plugging in $\nu = T^{1/3}$, we get the desired result. We are only left with proving Lemma 5 to complete the proof of Proposition 3.

Proof of Lemma 5: Recall from Section IV that policy π_D minimizes the bound on drift-plus-optimistic-penalty $\hat{L}^{\pi_D}(t)$. Therefore, by comparing π_D's drift-plus-optimistic-penalty $\hat{L}^{\pi_D}(t)$ against $\{\mu_{ij}^{stat}\}_{(i,j) \in \mathcal{E}}$, the solution to $\mathcal{P}$, we get

$$\mathbb{E}_{Q,\hat{c}}\Big[\Delta\Phi^{\pi_D}(t) + \nu \sum_{(i,j) \in \mathcal{E}} \mu_{ij}^{\pi_D}(t) \hat{c}_{ij}(t)\Big] \leq B + \lambda Q_s^{\pi_D}(t) + \sum_{ij \in \mathcal{E}} \mu_{ij}^{stat}(Q_j^{\pi_D}(t) - Q_i^{\pi_D}(t) + \nu \hat{c}_{ij}(t))$$
$$= B + \nu \sum_{(i,j) \in \mathcal{E}} \mu_{ij}^{stat} \hat{c}_{ij}(t) + Q_s^{\pi_D}(t)\Big[\lambda - \sum_{j \in \mathcal{N}_s} \mu_{sj}^{stat}\Big] + \sum_{i \in \mathcal{N} \setminus \{s\}} Q_i^{\pi_D}(t)\Big[\sum_{j : i \in \mathcal{N}_j} \mu_{ji}^{stat} - \sum_{j \in \mathcal{N}_i} \mu_{ij}^{stat}\Big].$$

The equality above is obtained by rearranging the summation. From the constraints of $\mathcal{P}$, we have $\lambda \leq \sum_{j \in \mathcal{N}_s} \mu_{sj}^{stat}$ and $\sum_{j : i \in \mathcal{N}_j} \mu_{ji}^{stat} \leq \sum_{j \in \mathcal{N}_i} \mu_{ij}^{stat}$, $\forall i \in \mathcal{N} \setminus \{s, d\}$. Using these in the above inequality concludes the proof of lemma. □

Appendix F
Proof of Proposition 4

Proposition 4: $R_4^{\pi_D}(T) = O(T^{2/3})$ with $\nu = T^{1/3}$.

Proof: Using Cauchy–Schwarz and Jensen's inequalities, we have $\sum_{i \in \mathcal{N}} \mathbb{E}[Q^{\pi_D}(T)^2] \geq \frac{1}{|\mathcal{N}|}(\sum_{i \in \mathcal{N}} \mathbb{E}[Q_i^{\pi_D}(T)])^2$. Combining this with the definition of Lyapunov function $\mathbb{E}[\Phi^{\pi_D}(T)] = \frac{1}{2}\sum_{i \in \mathcal{N}} \mathbb{E}[Q^{\pi_D}(T)^2]$, we have

$$R_4^{\pi_D}(T) = C_B \sum_{i \in \mathcal{N}} \mathbb{E}[Q_i^{\pi_D}(T)] \leq C_B \sqrt{2|\mathcal{N}|\, \mathbb{E}[\Phi^{\pi_D}(T)]}.$$

Lemma 6: $\mathbb{E}[\Phi^{\pi_D}(T)] = O(T^{4/3})$ with $\nu = T^{1/3}$.

From the lemma, we can get the desired result as $R_4^{\pi_D}(T) \leq C_B \sqrt{2|\mathcal{N}|\, \mathbb{E}[\Phi^{\pi_D}(T)]} = O(T^{2/3})$. We are left with proving Lemma 6 to conclude the proof of Proposition 4.

Proof of Lemma 6: To prove this lemma, we use the result from Lemma 5 from Proposition 3's proof. Taking expectation $\mathbb{E}[\cdot]$ on both sides of Lemma 5's result, using total law of expectations, rearranging, and summing over $t = 1, 2, ..., T-1$, we obtain $\mathbb{E}[\Phi^{\pi_D}(T)] \leq B + \nu \sum_t \sum_{ij \in \mathcal{E}} \mathbb{E}[\mu_{ij}^{stat} \hat{c}_{ij}(t)] - \nu \sum_t \sum_{ij \in \mathcal{E}} \mathbb{E}[\mu_{ij}^{\pi_D}(t) \hat{c}_{ij}(t)]$.

Since $\mathbb{E}[\hat{c}_{ij}(t)] = \mathbb{E}[\bar{c}_{ij}(t) - \sqrt{\beta \log(t/\delta)/N_{ij}(t)}] \leq C_{max}$ and $0 \leq \mu_{ij}^{stat} \leq \mu_{ij}^{max}$, we have

$$\sum_{t=1}^{T-1} \sum_{(i,j) \in \mathcal{E}} \mathbb{E}[\mu_{ij}^{stat} \hat{c}_{ij}(t)] \leq T \sum_{(i,j) \in \mathcal{E}} \mu_{ij}^{max} C_{max} = O(T).$$

Further, we have $\hat{c}_{ij}(t) = \bar{c}_{ij}(t) - \sqrt{\beta \log(t/\delta)/N_{ij}(t)} \geq -\sqrt{\beta \log(t/\delta)/N_{ij}(t)} \geq -\sqrt{\beta \log(T/\delta)/N_{ij}(t)}$. Therefore,

$$\sum_{t=1}^{T-1} \sum_{ij \in \mathcal{E}} \mathbb{E}[\mu_{ij}^{\pi_D}(t) \hat{c}_{ij}(t)] \geq -\sum_{t=1}^{T-1} \sum_{ij \in \mathcal{E}} \mathbb{E}\Big[\mu_{ij}^{\pi_D}(t) \sqrt{\frac{\beta \log \frac{T}{\delta}}{N_{ij}(t)}}\Big] \geq -O(\sqrt{T} \log T) \geq -O(T)$$

where, the second inequality above follows from Lemma 4 in Proposition 1's proof. Finally, plugging these back in the drift bound and taking $\nu = T^{1/3}$, we get $\mathbb{E}[\Phi^{\pi_D}(T)] = O(\nu T) = O(T^{4/3})$. This concludes the proof of lemma. □

REFERENCES

[1] Lin Xiao, M. Johansson and S. P. Boyd, "Simultaneous routing and resource allocation via dual decomposition," in IEEE Transactions on Communications, vol. 52, no. 7, pp. 1136-1144, July 2004, doi: 10.1109/TCOMM.2004.831346.

[2] R. L. Cruz and A. V. Santhanam, "Optimal routing, link scheduling and power control in multihop wireless networks," IEEE INFOCOM 2003. Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE Cat. No.03CH37428), San Francisco, CA, USA, 2003, pp. 702-711 vol.1, doi: 10.1109/INFCOM.2003.1208720.

[3] W. Fang, X. Liang, S. Li, L. Chiaraviglio, and N. Xiong, "VMPlanner: Optimizing virtual machine placement and traffic flow routing to reduce network power costs in cloud data centers," Computer Networks, vol. 57, no. 1. Elsevier BV, pp. 179–196, Jan. 2013. doi: 10.1016/j.comnet.2012.09.008.

[4] J. Ghaderi, S. Shakkottai, and R. Srikant, "Scheduling Storms and Streams in the Cloud," ACM Transactions on Modeling and Performance Evaluation of Computing Systems, vol. 1, no. 4. Association for Computing Machinery (ACM), pp. 1–28, Aug. 02, 2016. doi: 10.1145/2904080.

[5] E. Uysal-Biyikoglu, B. Prabhakar, and A. El Gamal, "Energy-efficient packet transmission over a wireless link," IEEE/ACM Transactions on Networking, vol. 10, no. 4. Institute of Electrical and Electronics Engineers (IEEE), pp. 487–499, Aug. 2002. doi: 10.1109/tnet.2002.801419.

[6] A. Fu, E. Modiano and J. Tsitsiklis, "Optimal energy allocation for delay-constrained data transmission over a time-varying channel," IEEE INFOCOM 2003. Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE Cat. No.03CH37428), San Francisco, CA, USA, 2003, pp. 1095-1105 vol.2, doi: 10.1109/INFCOM.2003.1208946.

[7] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," in IEEE Transactions on Automatic Control, vol. 37, no. 12, pp. 1936-1948, Dec. 1992, doi: 10.1109/9.182479.

[8] M. J. Neely, "Energy optimal control for time-varying wireless networks," in IEEE Transactions on Information Theory, vol. 52, no. 7, pp. 2915-2934, July 2006, doi: 10.1109/TIT.2006.876219.

[9] R. Urgaonkar, B. Urgaonkar, M. J. Neely, and A. Sivasubramaniam, "Optimal power cost management using stored energy in data centers," Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems. ACM, Jun. 07, 2011. doi: 10.1145/1993744.1993766.

[10] X. Qiu, H. Li, C. Wu, Z. Li and F. C. M. Lau, "Cost-Minimizing Dynamic Migration of Content Distribution Services into Hybrid Clouds," in IEEE Transactions on Parallel and Distributed Systems, vol. 26, no. 12, pp. 3330-3345, 1 Dec. 2015, doi: 10.1109/TPDS.2014.2371831.

[11] M. Gatzianas, L. Georgiadis and L. Tassiulas, "Control of wireless networks with rechargeable batteries [transactions papers]," in IEEE Transactions on Wireless Communications, vol. 9, no. 2, pp. 581-593, February 2010, doi: 10.1109/TWC.2010.080903.

[12] L. Ying, S. Shakkottai, A. Reddy, and S. Liu, "On Combining Shortest-Path and Back-Pressure Routing Over Multihop Wireless Networks," IEEE/ACM Transactions on Networking, vol. 19, no. 3. Institute of Electrical and Electronics Engineers (IEEE), pp. 841–854, Jun. 2011. doi: 10.1109/tnet.2010.2094204.

[13] J. Fu, B. Moran, J. Guo, E. W. M. Wong and M. Zukerman, "Asymptotically Optimal Job Assignment for Energy-Efficient Processor-Sharing Server Farms," in IEEE Journal on Selected Areas in Communications, vol. 34, no. 12, pp. 4008-4023, Dec. 2016, doi: 10.1109/JSAC.2016.2611864.

[14] Z. Zhang, M. Zhang, A. Greenberg, Y. C. Hu, R. Mahajan, and B. Christian, "Optimizing cost and performance in online service provider networks," in Proc. Usenix Symp. Networked Syst. Des. Implement., San Jose, CA, USA, Apr. 2010, pp. 33–48.

[15] H. Liao et al., "Learning-Based Context-Aware Resource Allocation for Edge-Computing-Empowered Industrial IoT," in IEEE Internet of Things Journal, vol. 7, no. 5, pp. 4260-4277, May 2020, doi: 10.1109/JIOT.2019.2963371.

[16] Z. Guo et al., "AggreFlow: Achieving Power Efficiency, Load Balancing, and Quality of Service in Data Center Networks," in IEEE/ACM Transactions on Networking, vol. 29, no. 1, pp. 17-33, Feb. 2021, doi: 10.1109/TNET.2020.3026015.

[17] T. L. Lai, and H. Robbins, "Asymptotically efficient adaptive allocation rules," Advances in applied mathematics, 6(1), 4-22, 1985.

[18] P. Auer, "Using confidence bounds for exploitation-exploration tradeoffs," Journal of Machine Learning Research, 397-422, 2002.

[19] S. Bubeck and N. Cesa-Bianchi, "Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems," arXiv, Nov. 03, 2012. Accessed: Nov. 14, 2023. [Online]. Available: http://arxiv.org/abs/1204.5721

[20] B. Awerbuch and R. Kleinberg, "Online linear optimization and adaptive routing," Journal of Computer and System Sciences, vol. 74, no. 1, pp. 97–114, Feb. 2008, doi: 10.1016/j.jcss.2007.04.016.

[21] V. Dani, T. P. Hayes, and S. M. Kakade, "Stochastic Linear Optimization Under Bandit Feedback," COLT. Vol. 2. 2008.

[22] P. Rusmevichientong and J. N. Tsitsiklis, "Linearly Parameterized Bandits," Mathematics of Operations Research, vol. 35, no. 2. Institute for Operations Research and the Management Sciences (INFORMS), pp. 395–411, May 2010. doi: 10.1287/moor.1100.0446.

[23] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," Advances in neural information processing systems, 24, 2011.

[24] X. Fu and E. Modiano, "Optimal Routing to Parallel Servers With Unknown Utilities—Multi-Armed Bandit With Queues," IEEE/ACM Trans. Networking, pp. 1–16, 2022, doi: 10.1109/TNET.2022.3227136.

[25] O. Amar, I. Sarfati and K. Cohen, "An Online Learning Approach to Shortest Path and Backpressure Routing in Wireless Networks," in IEEE Access, vol. 11, pp. 57253-57267, 2023, doi: 10.1109/ACCESS.2023.3282365.

[26] M. J. Neely, "Stochastic Network Optimization with Application to Communication and Queueing Systems." Springer International Publishing, 2010. doi: 10.1007/978-3-031-79995-2.

[27] 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. Now Publishers, pp. 1–144, 2005. doi: 10.1561/1300000001.

[28] M. J. Wainwright, "High-Dimensional Statistics." Cambridge University Press, Feb. 12, 2019. doi: 10.1017/9781108627771.

[29] L. Besson and E. Kaufmann, "What Doubling Tricks Can and Can't Do for Multi-Armed Bandits." arXiv, 2018. doi: 10.48550/arxiv.1803.06971.