Learning While Scheduling in Multi-Server Systems With Unknown Statistics: MaxWeight with Discounted UCB

Zixian Yang
University of Michigan

R. Srikant
University of Illinois at Urbana-Champaign

Lei Ying
University of Michigan

Abstract

Multi-server queueing systems are widely used models for job scheduling in machine learning, wireless networks, and crowdsourcing. This paper considers a multi-server system with multiple servers and multiple types of jobs, where different job types require different amounts of processing time at different servers. The goal is to schedule jobs on servers without knowing the statistics of the processing times. To fully utilize the processing power of the servers, it is known that one has to at least learn the service rates of different job types on different servers. Prior works on this topic decouple the learning and scheduling phases which leads to either excessive exploration or extremely large job delays. We propose a new algorithm, which combines the MaxWeight scheduling policy with discounted upper confidence bound (UCB), to simultaneously learn the statistics and schedule jobs to servers. We obtain performance bounds for our algorithm that hold for both stationary and nonstationary service rates. Simulations confirm that the delay performance of our algorithm is several orders of magnitude better than previously proposed algorithms. Our algorithm also has the added benefit that it can handle non-stationarity in the service processes.

1 INTRODUCTION

A multi-server system is a system with multiple servers for serving jobs of different types as shown in Figure 1. An incoming job can be served by one of the servers and the service time depends on both the server and the job type. Multi-server systems have been used to model many real-world applications in communication and computer systems such as load balancing in a cloud-computing cluster, packet scheduling in multi-channel wireless networks, crowdsourcing, etc. In cloud-computing, a job may be a machine learning task and a server may be a virtual machine or a container, so the processing time of the machine learning task depends on the virtual machine's configuration. In crowdsourcing, jobs could be tagging of images and servers are workers, so the amount of the time a worker takes to tag the images depends on her familiarity of the images. In both cases, the scheduler may not know the statistics of process times of a server before a sufficient number of jobs of the same type are processed at the server.

When the mean server times are known, the best known algorithm for scheduling in multi-server systems is the celebrated MaxWeight algorithm proposed by Tassiulas and Ephremides (1993). Let $Q_i(t)$ denote the number of type-i jobs waiting to be served and $1/\mu_{i,j}$ denote the mean service time of serving a type-i job at server j. When server j is available, MaxWeight schedules a type i_j^* job to server j such that

$$i_j^* \in \arg\max Q_i(t)\mu_{i,j}.$$

A set of arrival rates is said to be supportable if there exists a scheduling algorithm such that, under this set of arrival rates, the queue lengths are bounded in an appropriate sense. The MaxWeight algorithm is provably throughput optimal (Tassiulas and Ephremides, 1993), i.e., it has the largest set of supportable arrival rates, also called the capacity region. Besides throughput optimality, MaxWeight has also near-optimal delay performance in various settings (Stolyar, 2004; Andrews et al., 2007; Shah and Wischik, 2007; Kang and Williams, 2013; Eryilmaz and Srikant, 2012; Maguluri and Srikant, 2016).

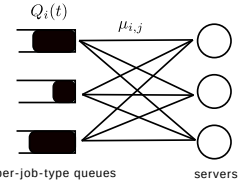


Figure 1: A Multi-Server System With Three Servers and Three Types of Jobs.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

A key assumption behind the MaxWeight algorithm is that the scheduler knows the mean service rates $\mu_{i,j}$ for all i and j. This assumption is becoming increasingly problematic in emerging applications such as cloud computing and crowdsourcing due to either high variability of jobs (such as complex machine learning tasks) or servers (such as human experts in crowdsourcing). In these emerging applications, the mean service rates need to be learned while making scheduling decisions. Therefore, learning and scheduling are coupled and jointly determine the performance of the system.

A straightforward idea to learn the mean service rates is using the sample average, i.e., replacing $\mu_{i,j}$ with $1/\bar{s}_{i,j}$ where $\bar{s}_{i,j}$ is the empirical mean of the service time of type-i jobs at server j, based on the jobs completed at server j so far. However, because of the coupling between learning and scheduling, this approach can be unstable. We have not seen an explicit counter-example in the published literature which shows such instability can occur, so we provide one in Appendix A. From the example, we observe that the problem of using empirical mean is that the initial bad samples led to a poor estimation of $\mu_{i,j}$, which led to poor scheduling decisions. They stop the scheduler from getting new samples from other queue-server pairs and therefore the system is "locked in" in a state with poor estimation and wrong scheduling decisions, which led to instability.

To overcome this problem, as in multi-armed bandit problems, we should encourage exploration: since the service rate of a server for a particular job can be estimated only by repeatedly scheduling jobs on all jobs, we should occasionally schedule jobs even on servers whose service rates are estimated to be small to overcome poor estimates due to randomness. For example, as in online learning, we can add an exploration bonus $b_{i,j}$, e.g., the upper confidence bound (UCB), to the empirical mean $\hat{\mu}_{i,j}$. Indeed, there have been a sequence of recent studies that study job scheduling in multi-server systems as an online learning problem (multi-armed bandits or linear bandits). We now review different categories of prior work along these lines and place our work in the context of the prior work:

Queue-blind Algorithms: Queue blind algorithms do not take queue lengths into consideration at all when making scheduling decisions. In one line of work, the performance metric is the total reward received from serving jobs (Li et al., 2019; Liu et al., 2021); however, for such algorithms, the queue lengths can potentially blow up to infinity asymptotically, which means that finite-time bounds for queue lengths can be excessively large and thus, such algorithms cannot be used in practice. Another line of work in the context of queue-blind scheduling algorithms addresses stability by assuming that the arrival rates of each type of job is known. They then use well-known scheduling algorithms such as $c\mu$-rule (Krishnasamy et al., 2018b) or weighted random routing (Choudhury et al., 2021) or utility-based joint learning and scheduling (Hsu et al., 2022). The drawback of such algorithms is that queue lengths can still be excessive large even if the queue lengths do not blow up to infinity asymptotically. The reason is the knowledge of queue lengths can encourage a phenomenon called *resource pooling* which leads to greater efficiency. While we will not spend too much space explaining the concept of resource pooling, we hope that the following example clarifies the situation. Suppose you visit a grocery store and are not allowed to look at the queue lengths at each checkout lane before joining the checkout line. Then, some checkout lines can be excessively long, while others may even be totally empty. On the other hand, in practice, we look at the length of each checkout line and join the shortest one, which results in much better delay performance.

Queue-Aware Algorithms: In early work on the problem (Neely et al., 2012; Krishnasamy et al., 2018a, 2021; Yekkehkhany and Nagi, 2020), a fraction of time is allocated to probing the servers and the rest of the time is used to exploit this information. In the context of our problem, we would end up exploring all (job type, server) pairs the same number of times which is wasteful. On the other hand, exploration and exploitation are decoupled in such a forced exploration, which makes it easier to derive analytical derivation of performance bounds . If one uses optimistic exploration such as UCB or related algorithms, the queue length information and the UCB-style estimation are coupled, which makes it difficult to analyze the system. Two approaches to decoupling UCB-style estimators have been studied prior to our paper: (a) In the work by Stahlbuhk et al. (2019), the algorithm proceeds in frames (a frame is a collection of contiguous time slots), where the queue length information is frozen at the beginning of each frame and UCB is used to estimate the service rates of the servers; additionally, UCB is reset at the end of each frame, and (b) In the work by Freund et al. (2022), a schedule is fixed throughout each phase and thus, UCB is only executed for the jobs which are scheduled in that frame. The correlation between queues and UCB is more complicated here than in the algorithm of Stahlbuhk et al. (2019), which requires more sophisticated analysis to conclude stability. Our paper does not explicitly decouple exploration and exploitation. In fact, we continuously update the UCB bonuses and perform scheduling at each time instant but we use a version of UCB tailored to nonstationary environments. This allows our algorithm to quickly adapt to changes even in stationary settings, in addition to having the advantage of being able to handle nonstationary environments. On the other hand, the fact that the schedule and UCB are updated at every time step means that we require a new analysis of stability. In particular, unlike prior work, our approach requires the use of concentration results for self-normalized means from Garivier and Moulines (2008). In addition to differences in the algorithms and analysis, we also note other key differences between our paper and theirs (Stahlbuhk et al., 2019; Freund

et al., 2022): Stahlbuhk et al. (2019) considers scheduling in a general conflict graph, which includes our multi-server model as a special case. Freund et al. (2022) considers a general multi-agent setting that includes the centralized case as a special case. Both Stahlbuhk et al. (2019) and Freund et al. (2022) assume the system is stationary but Freund et al. (2022) allows dynamic arrivals and departures of queues while our paper studies a nonstationary, centralized setting that includes the stationary setting with a fixed set of queues as a special case.

The main contributions of this paper are summarized below.

- **Theoretical Results:** We introduce MaxWeight with discounted UCB in this paper. Discounted UCB was first proposed for nonstationary bandit problems (Kocsis and Szepesvári, 2006). For our problem, with a revised discounted UCB, the values of UCB bonuses depend on limited past history, instead of the entire history, which allows us to handle the coupling between the queue lengths and UCB bonuses. We establish the queue stability of MaxWeight with discounted UCB for nonstationary environments where the arrival rates and service rates may change over time. Given that the variation of service rates during the service time of a single job is bounded by d ($d \leqslant 1$), we show that MaxWeight with discounted UCB can support any arrival rate vector $\boldsymbol{\lambda}$ such that $\boldsymbol{\lambda}+\delta\mathbf{1}$ is in the capacity region for some $\delta = \tilde{\Theta}(d)$, and the asymptotic time average of the expected queue length is bounded by $\tilde{O}(1/\delta_{\max}^3)$, where $\delta_{\max}$ is the largest δ such that $\boldsymbol{\lambda} + \delta\mathbf{1}$ is in the capacity region. This queue length bound holds for both stationary and nonstationary settings and matches the bound of the stationary setting in (Freund et al., 2022) in terms of the traffic slackness $\delta_{\max}$.

- **Methodology:** Our analysis is based on Lyapunov drift analysis. However, there are several difficulties due to joint scheduling and learning. For example, the estimated mean service time is the discounted sum of previous service times divided by the sum of the discount coefficients and the summation is taken over the time slots in which there is job completion, which themselves are random variables depending on the scheduling and learning algorithm. To deal with this difficulty, we first transform the summation into a summation over the time slots in which a job starts, and then use a Hoeffding-type inequality for *self-normalized means* with a random number of summands (Garivier and Moulines, 2008, Theorem 18)(Garivier and Moulines, 2011) to obtain a concentration bound. Another difficulty is in bounding the discounted number of times server j serves type-i jobs. Our method is to divide the interval into sub-intervals of carefully chosen lengths so that the discount coefficients can be lower bounded by a constant in each sub-interval. We believe these ideas may be useful for analyzing other joint learning and scheduling algorithms as well.

- **Numerical Studies:** We compare the proposed algorithm with previously proposed algorithms in the literature. The results show that our algorithms achieves delays that are several orders of magnitude smaller than previously proposed algorithms. A noteworthy observation is that, the discounted UCB algorithm which was originally designed for nonstationary environments, allows us to design a joint learning/scheduling algorithm which outperforms the state-of-the art even in stationary environments, as shown in Fig. 2. We believe the main reason is that we do not decouple learning and scheduling explicitly and the resulting continuous updates to the learning and scheduling decisions are essential to achieve good delay performance.

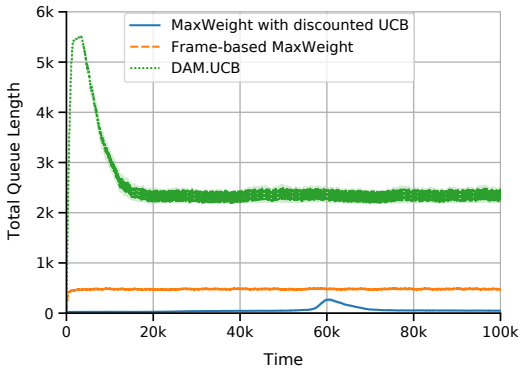


Figure 2: Comparison Among MaxWeight With Discounted UCB (Proposed), Frame-Based MaxWeight (Stahlbuhk et al., 2019), and DAM.UCB (Freund et al., 2022).

2 MODEL

We consider a multi-server system with J servers, indexed with $j \in \{1, 2, \ldots, J\}$, and I types of jobs, indexed with $i \in \{1, 2, \ldots, I\}$. The system maintains a separate queue for each job type, as shown in Figure 1.

We consider a discrete-time system. The number of jobs that arrive at queue i is denoted by $(A_i(t))_{t\geqslant 0}$ where t denotes the time slot. Assume that $(A_i(t))_{t\geqslant 0}$ are independent with unknown mean $E[A_i(t)] = \lambda_i(t)$ and are bounded, i.e., $A_i(t) \leqslant U_{\mathrm{A}}$ for all i and t. Let $\boldsymbol{A}(t) := (A_i(t))_{i=1}^I$ and $\boldsymbol{\lambda}(t) := (\lambda_i(t))_{i=1}^I$.

We say a server is available in time slot t if the server is not serving any job at the beginning of time slot t; otherwise, we say the server is busy. At the beginning of each time slot, each available server picks a job from one of the queues. Note that each server can serve at most one job at a time and can start to serve another job only after finishing the current job, i.e., the job scheduling is non-preemptive. When a job from queue i (job of type i) is picked by server j in time slot t, it requires $S_{i,j}(t)$ time slots to finish serving the job. For any i, j, $(S_{i,j}(t))_{t\geqslant 0}$

are independent random variables with unknown mean $E[S_{i,j}(t)] = \frac{1}{\mu_{i,j}(t)}$ and are bounded, i.e., $S_{i,j}(t) \leqslant U_{\mathrm{S}}$ for all i, j and t. $A_i(t)$ and $S_{i,j}(t)$ for different i, j are also independent. Let $\boldsymbol{S}(t) := (S_{i,j}(t))_{i=1,\ldots,I,j=1,\ldots,J}$ and $\boldsymbol{\mu}(t) := (\mu_{i,j}(t))_{i=1,\ldots,I,j=1,\ldots,J}$. Note that we allow $\boldsymbol{\lambda}(t)$ and $\boldsymbol{\mu}(t)$ to be time-varying to model nonstationary environments, and the value of $S_{i,j}(t)$ is generated at time slot t and will not change after that.

If server j is available and picks queue i in time slot t or if server j is busy serving queue i in time slot t, we say server j is scheduled to queue i in time slot t. Let $I_j(t)$ denote the queue to which server j is scheduled in time slot t. Define a waiting queue $\tilde{Q}_i(t)$ for each job type i. A job of type i joins the waiting queue $\tilde{Q}_i(t)$ when it arrives, and leaves the waiting queue $\tilde{Q}_i(t)$ when it is picked by a server under the algorithm. If an available server j picks queue i in time slot t and there is no job in the waiting queue i, i.e., $\tilde{Q}_i(t) + A_i(t) = 0$, we say server j is idling in time slot t and the server j will be available in the next time slot. Let $\eta_j(t)$ be an indicator function such that $\eta_j(t) = 1$ if server j is not idling in time slot t and $\eta_j(t) = 0$ otherwise. Let $\mathbb{1}_{i,j}(t)$ be another indicator function such that $\mathbb{1}_{i,j}(t) = 1$ if $I_j(t) = i$ and server j finishes serving the job of type i at the end of time slot t, or if $I_j(t) = i$ and server j is idling.

Let $Q_i(t)$ denote the actual queue length of jobs at queue i at the beginning of time slot t so $Q_i(t)$ is the total number of type-i jobs in the system. Thus, $\tilde{Q}_i(t)$ is $Q_i(t)$ minus the number of type i jobs that are in service. A job leaves the actual queue only when it is completed. Then we have the following queue dynamics:

$$Q_i(t+1) = Q_i(t) + A_i(t) - \sum_j \mathbb{1}_{i,j}(t)\eta_j(t). \quad (1)$$

Our objective is to find an efficient learning and scheduling algorithm to stabilize $Q_i(t)$ for all i, i.e., preventing the queue lengths from going to infinity. In each time slot, the scheduling algorithm decides which queue to serve for each available server.

3 ALGORITHM — MAXWEIGHT WITH DISCOUNTED-UCB

We propose MaxWeight with Discounted-UCB algorithm, which combines the MaxWeight scheduling algorithm (Tassiulas and Ephremides, 1992) with discounted UCB (Kocsis and Szepesvári, 2006) for learning the service statistics, as shown in Algorithm 1.

In Algorithm 1, we first fix the discount factor γ beforehand and initialize the estimates $\hat{N}_{i,j}(0)$, $\hat{\phi}_{i,j}(0)$, and the counter $M_{i,j}(0)$, as shown in Line 1. In the algorithm, $\hat{N}_{i,j}(t)$ is the discounted number of type-i jobs served by server j by time slot t and $\hat{\phi}_{i,j}(t)$ is the discounted number of time slots used by server j for serving type-i jobs by time slot t. If server j is currently serving a type-i job, $M_{i,j}(t)$ is the service time the job has received by time slot t (not including time slot t); otherwise, $M_{i,j}(t) = 0$. Next, we define $g(\gamma)$ in Line 2. Note that $g(\gamma)$ can be easily computed using numerical methods. For intuition, $g(\gamma) \approx \frac{8}{1-\gamma}$ and larger γ implies a larger $g(\gamma)$. At time $t = 0$, we schedule each server to the queues uniformly at random. If $t \geqslant 1$, we first update our estimates of service rates and the UCB bonuses and then do the scheduling using the MaxWeight algorithm with the true service rates replaced by the UCB. Specifically, at the beginning of each time slot t, we update $\hat{N}_{i,j}(t)$ and $\hat{\phi}_{i,j}(t)$ as follows:

$$\begin{aligned}\hat{N}_{i,j}(t) =& \gamma\hat{N}_{i,j}(t-1) + \gamma^{M_{i,j}(t-1)}\mathbb{1}_{i,j}(t-1)\eta_j(t-1)\\ \hat{\phi}_{i,j}(t) =& \gamma\hat{\phi}_{i,j}(t-1)\\ &+ \gamma^{M_{i,j}(t-1)}\mathbb{1}_{i,j}(t-1)\eta_j(t-1)M_{i,j}(t). \quad (2)\end{aligned}$$

That is, if the job has not yet finished or the server is idling, we simply multiply $\hat{N}_{i,j}(t-1)$ and $\hat{\phi}_{i,j}(t-1)$ by a discount factor γ; if the server is not idling and the job has finished, we update $\hat{N}_{i,j}(t-1)$ by multiplying γ and adding a number $\gamma^{M_{i,j}(t-1)}$ and update $\hat{\phi}_{i,j}(t-1)$ by multiplying γ and adding a discounted service time. The discount $\gamma^{M_{i,j}(t-1)}$

Algorithm 1: MaxWeight With Discounted-UCB

1: **Initialize:** Fix $\gamma \in (0,1)$; $\hat{N}_{i,j}(0) = 0$, $\hat{\phi}_{i,j}(0) = 0$, $M_{i,j}(0) = 0$ for all i, j.
2: Define $g(\gamma)$ such that $\gamma = 1 - \frac{8\log g(\gamma)}{g(\gamma)}$.
3: If $t = 0$, schedule each server to the queues uniformly at random.
4: **for** $t = 1$ to infinity **do**
5: **for** $i = 1, \ldots, I$ and $j = 1, \ldots, J$ **do**
6: **if** $I_j(t-1) = i$ **then**
7: $M_{i,j}(t) = M_{i,j}(t-1) + 1$ // the number of time slots already served
8: **end if**
9: Update $\hat{N}_{i,j}(t)$ and $\hat{\phi}_{i,j}(t)$ according to (2).
10: $\hat{\mu}_{i,j}(t) = \frac{\hat{N}_{i,j}(t)}{\hat{\phi}_{i,j}(t)}$ // estimate of the service rate
11: $b_{i,j}(t) = \min\left\{c_1 U_{\mathrm{S}}\sqrt{\frac{\log g(\gamma)}{\hat{N}_{i,j}(t)}}, 1\right\}$ // UCB bonus term, where $c_1 > 0$ is a constant
12: **if** $\mathbb{1}_{i,j}(t-1) = 1$ **then**
13: $M_{i,j}(t) = 0$ // reset the counter if the server becomes available.
14: **end if**
15: **end for**
16: **for** $j = 1, \ldots, J$ **do**
17: **if** server j is available **then**
18: $\hat{i}^*_j(t) = \arg\max_i Q_i(t)\left(\hat{\mu}_{i,j}(t) + b_{i,j}(t)\right)$ // server j picks $\hat{i}^*_j(t)$
19: **end if**
20: **end for**
21: **end for**

actually means that the service time is discounted starting from the time when the job starts. This update is slightly different from the discounted UCB in (Kocsis and Szepesvári, 2006) and is needed for a technical reason. Then we obtain $\hat{\mu}_{i,j}(t)$, an estimate of the service rate, as shown in Line 10, where we use the convention that $0/0 = 0$. For each available server, we pick the queue with the largest product of queue length and UCB of the service rate, as shown in Line 18, where $\hat{i}_j^*(t)$ denotes the queue that server j picks and ties are broken arbitrary.

In a stationary environment, the use of discounted average instead of simple average reduces the influence of previous service times on the current estimate, and weakens the dependence between queue lengths and UCB bonuses. In a nonstationary environment, it ensures that the estimation process can adapt to the nonstationary service rate since the discount factor reduces the influence of previous service times on the current estimate. UCB helps with the exploration of the service times for different servers and job types. The MaxWeight algorithm is known to be throughput optimal (Srikant and Ying, 2014). These ideas are combined in the proposed algorithm.

4 MAIN RESULT

In this section, we will present our main result. We consider Algorithm 1 with a sufficiently large γ such that $g(\gamma) \geqslant \max\{e^5, 8U_{\mathrm{S}}\}$. We make the following assumption on the time-varying mean service times and rates:

Assumption 1. *$\mu_{i,j}(t)$ satisfies the following two conditions:*

(1) For any i, j and any t_a, t_b such that $t_a \neq t_b$ and $|t_a - t_b| \leqslant 2g(\gamma)$,

$$\left|\frac{1}{\mu_{i,j}(t_a)} - \frac{1}{\mu_{i,j}(t_b)}\right| \leqslant \frac{1}{g(\gamma)}\left(\frac{1}{\gamma}\right)^{|t_a - t_b|-1};$$

(2) There exists an absolute constant $p > 0$ such that for any i, j and any t_a, t_b such that $|t_a - t_b| \leqslant U_{\mathrm{S}}$,

$$|\mu_{i,j}(t_a) - \mu_{i,j}(t_b)| \leqslant \frac{1}{[g(\gamma)]^p}.$$

Remark 1. *Note that in the first condition in Assumption 1, $\frac{1}{\gamma} > 1$, so the allowable change of the mean service time increases exponentially with respect to the time difference. Therefore, the second condition in Assumption 1 will be dominating for large $|t_a - t_b|$. Recall that $g(\gamma) \approx \frac{8}{1-\gamma}$, so the bound in condition (2) is roughly equivalent to that the maximum change that can occur when serving a job is $\frac{(1-\gamma)^p}{8^p}$ for some $p > 0$ (note that U_{S} is an upper bound on the service times). This bound increases as γ decreases because the algorithm can quickly adapt by aggressively discounting the past samples.*

For the nonstationary system considered in this paper, we introduce the following definition $\mathcal{C}(W)$ for the capacity region:

$$\begin{aligned}\mathcal{C}(W) = \Big\{ &(\boldsymbol{R}(t))_{t\geqslant 0} : \text{there exists } (\boldsymbol{\alpha}(t))_{t\geqslant 0} \text{ such that} \\ &\sum_i \alpha_{i,j}(t) \leqslant 1 \text{ for all } j, t \text{ and for any } i, t, \\ &\text{there exists } w(t) \text{ such that } 1 \leqslant w(t) \leqslant W \text{ and} \\ &\sum_{\tau=t}^{t+w(t)-1} R_i(\tau) \leqslant \sum_{\tau=t}^{t+w(t)-1}\sum_j \alpha_{i,j}(\tau)\mu_{i,j}(\tau)\Big\},\end{aligned} \tag{3}$$

where $\boldsymbol{\alpha}(t) := (\alpha_{i,j}(t))_{i=1,\ldots,I,j=1,\ldots,J}$ and $W \geqslant 1$ is a constant. $\boldsymbol{R}(t)$ can be interpreted as allocatable service rates for time t. This capacity region means that for some $(\boldsymbol{R}(t))_{t\geqslant 0}$ in this region, for any time t and queue i, there exists a time window such that the sum of $R_i(t)$ over this time window is less than the sum of appropriately allocated service rates. If $(\boldsymbol{\alpha}(t))_{t\geqslant 0}$ is given, then a randomized scheduling algorithm using $(\boldsymbol{\alpha}(t))_{t\geqslant 0}$ guarantees that the service rate received by queue i in a time window is at least as large as the sum of $R_i(t)$ in this time window. Note that $\mathcal{C}(W_1) \subseteq \mathcal{C}(W_2)$ if $W_1 \leqslant W_2$. If $W = 1$ and $\boldsymbol{R}(t)$ and $\boldsymbol{\mu}(t)$ are time-invariant, then this definition reduces to the capacity region definition for the stationary setting (Srikant and Ying, 2014). Let $\boldsymbol{\lambda} := (\boldsymbol{\lambda}(t))_{t\geqslant 0}$. We assume that the arrival rates satisfy that $\boldsymbol{\lambda} + \delta\mathbf{1} \in \mathcal{C}(W)$, where $\mathbf{1}$ denotes an all-ones vector and we assume that $W \leqslant \frac{g(\gamma)}{2}$. We present Theorem 1 which shows that the MaxWeight with Discounted-UCB algorithm can stabilize the queues with such arrival rates. Another interpretation is that our algorithm can stabilize any arrival rate that satisfies $\lambda_i(t) + \delta \leqslant R_i(t)$ for all i, t for some $(\boldsymbol{R}(t))_{t\geqslant 0}$ in the capacity region.

Theorem 1. *Consider Algorithm 1 with $c_1 = 4$ and $g(\gamma) \geqslant \max\{e^5, 8U_{\mathrm{S}}\}$. Suppose $Q_i(0) = 0$ for all i. Under Assumption 1, for arrival rates that satisfy $\boldsymbol{\lambda} + \delta\mathbf{1} \in \mathcal{C}(W)$, where $W \leqslant \frac{g(\gamma)}{2}$ and*

$$\delta \geqslant \frac{804IJU_{\mathrm{S}}^2 \log g(\gamma)}{[g(\gamma)]^{\min\{\frac{1}{2},p\}}}, \tag{4}$$

we have

$$\begin{aligned}\frac{1}{t}\sum_{\tau=1}^{t} E\left[\sum_i Q_i(\tau)\right] &\leqslant \frac{IU_{\mathrm{A}}g^2(\gamma)}{t} \\ &+ \left(1 + \frac{W}{t}\right)\left(\frac{1642I^2J^2U_{\mathrm{S}}^2U_{\mathrm{A}}^2 g(\gamma)}{\delta} + \frac{IU_{\mathrm{A}}^2 g^2(\gamma)}{\delta[t + 1 - g(\gamma)]}\right)\end{aligned}$$

for any $t \geqslant g(\gamma)$, and thus

$$\limsup_{t\to\infty} \frac{1}{t}\sum_{\tau=1}^{t} \mathbb{E}\left[\sum_i Q_i(\tau)\right] \leqslant \frac{1642I^2J^2U_{\mathrm{S}}^2U_{\mathrm{A}}^2 g(\gamma)}{\delta}.$$

We will discuss Theorem 1 in the stationary setting and the nonstationary setting in the following paragraphs. Note that the value of δ, the traffic slackness, measures the throughput loss, under MaxWeight with discounted UCB for given discount factor γ.

Stationary Setting For the stationary setting, Theorem 1 implies that if $g(\gamma)$ is sufficiently large, i.e., γ is sufficiently close to 1, then δ can be arbitrarily close to zero and hence the proposed algorithm can stabilize the queues with arrivals inside the capacity region, which means the throughput loss is close to zero. Given an arrival rate vector $\boldsymbol{\lambda}$ and letting $\delta_{\max}$ denote the largest δ such that $\boldsymbol{\lambda} + \delta\mathbf{1} \in \mathcal{C}(W)$, Theorem 1 implies that the asymptotic time average of expected queue length is bounded by $\tilde{O}(1/\delta_{\max}^3)$, which is obtained by setting $g(\gamma) = \tilde{\Theta}(1/\delta_{\max}^2)$ that satisfies the condition (4), where p can be set to an arbitrary large value because Assumption 1 always holds in the stationary setting.

Nonstationary Setting For the nonstationary setting, Assumption 1 comes into play because we need to consider the variation of service rates. Suppose that the variation of service rates within the service time of a single job is bounded by d. We want to obtain the smallest δ in Theorem 1, i.e., minimizing the throughput loss, while satisfying Assumption 1. We only consider the second condition in Assumption 1 since it is dominating as discussed in Remark 1. We consider the following two cases:

(A) For any $p \geqslant 1/2$, we choose $g(\gamma) = 1/d^{1/p}$. We can see that Assumption 1 (2) is satisfied with this p. Then By (4), δ can be as small as $\delta = 804IJU_{\mathrm{S}}^2 \log g(\gamma)/\sqrt{g(\gamma)}$. Note that $804IJU_{\mathrm{S}}^2 \log g(\gamma)/\sqrt{g(\gamma)}$ is decreasing in $g(\gamma)$ and $1/d^{1/p}$ is decreasing in p. Hence, we will choose $p = 1/2$ in order to obtain the smallest δ. Therefore, by setting $g(\gamma) = 1/d^2$, δ can be as small as $\delta = 804IJU_{\mathrm{S}}^2 d\log(1/d^2)$ in this case.

(B) For any $p < 1/2$, we choose $g(\gamma) = 1/d^{1/p}$. We can see that Assumption 1 (2) is satisfied. Then By (4), δ can be as small as $\delta = 804IJU_{\mathrm{S}}^2 d\log(1/d^{1/p})$, which is larger than that in Case (A) since $1/d^{1/p} > 1/d^2$.

Note that in each case although choosing $g(\gamma) < 1/d^{1/p}$ also satisfies Assumption 1 (2), it will induce a larger throughput loss since the right-hand side of (4) is decreasing in $g(\gamma)$. Combining these two cases, we conclude that if we choose $g(\gamma) = 1/d^2$, Assumption 1 (2) is satisfied with $p = 1/2$, and the smallest possible δ in Theorem 1 is of order $\tilde{\Theta}(d)$. In other words, the throughput loss is almost linear in terms of the variation d. Consider an arrival rate vector $\boldsymbol{\lambda}$ and let $\delta_{\max}$ denote the largest δ such that $\boldsymbol{\lambda} + \delta\mathbf{1} \in \mathcal{C}(W)$. Suppose $\delta_{\max}$ is greater than the smallest possible δ. Then $\delta_{\max} \geqslant \tilde{\Theta}(d)$. Hence, by setting $g(\gamma) = \tilde{\Theta}(1/\delta_{\max}^2)$, Assumption 1 (2) is satisfied with $p = 1/2$, and Theorem 1 implies that the asymptotic time average of expected queue length is bounded by $1642I^2J^2U_{\mathrm{S}}^2U_{\mathrm{A}}^2 g(\gamma)/\delta_{\max} = \tilde{O}(1/\delta_{\max}^3)$.

In many networks of interest, the arrival rates of flows are controlled by an algorithm called the congestion control protocol (Srikant and Ying, 2014). For congestion controlled flows, $\delta_{\max}$ is typically small; and for non-congestion controlled flows, called best-effort arrivals, $\delta_{\max}$ varies a lot.

We also want to point out that the assumption $W \leqslant g(\gamma)/2$ in Theorem 1 is reasonable. In fact, W captures the timescale at which congestion controlled arrivals react to nonstationarity. Recall from Assumption 1 that $1/g(\gamma)$ can loosely quantify the amount of nonstationarity the proposed algorithm can handle. Therefore, when the level of nonstationarity is high, the congestion controller needs to react faster, resulting in a small W.

5 PROOF ROADMAP

Our proof of Theorem 1 is based on Lyapunov drift analysis. Consider the Lyapunov function $L(t) \coloneqq \sum_i Q_i^2(t)$. We next present a proof roadmap. The complete proof of the theorem and the proofs of all the lemmas can be found in Appendix B and Appendix C.

5.1 Decomposing the Lyapunov Drift

First, we will divide the time horizon into intervals and later we can analyze the Lyapunov drift in each interval. Let $D_k(\gamma)$ denote the length of the k^{th} interval. The details of how we construct $D_k(\gamma)$ can be found in Appendix B.1. The main idea is that we want to make sure that $D_k(\gamma)$ is approximately $g(\gamma)$ so that the estimates of the mean service times in the current interval will "forget" the old samples in previous intervals due to the discount factor γ. Let $D_k \coloneqq D_k(\gamma)$ for ease of notation. Define $t_0 = 0$ and $t_k = t_{k-1} + D_{k-1}$ for $k \geqslant 1$. Then $[t_k, t_{k+1}]$ is the k^{th} interval.

Next, we analyze the Lyapunov drift in the $(k+1)^{\text{th}}$ interval given the queue length $\boldsymbol{Q}(t_k)$ and $\boldsymbol{H}(t_k)$ at the beginning of the k^{th} interval, where $\boldsymbol{Q}(t) \coloneqq (Q_i(t))_{i=1,\dots,I}$ and $\boldsymbol{H}(t)$ is defined as:

$$\boldsymbol{H}(t) \coloneqq \left(\tilde{\boldsymbol{Q}}(t), \boldsymbol{M}(t), \hat{\boldsymbol{N}}(t), \hat{\boldsymbol{\phi}}(t)\right),$$

where $\tilde{\boldsymbol{Q}}(t) \coloneqq (\tilde{Q}_i(t))_i$, $\boldsymbol{M}(t) \coloneqq (M_{i,j}(t))_{i,j}$, $\hat{\boldsymbol{N}}(t) \coloneqq (\hat{N}_{i,j}(t))_{i,j}$, and $\hat{\boldsymbol{\phi}}(t) \coloneqq (\hat{\phi}_{i,j}(t))_{i,j}$. Utilizing the queue dynamics (1), we can bound the Lyapunov drift by

$$\begin{aligned} &E\left[L(t_{k+1} + D_{k+1}) - L(t_{k+1}) \,|\boldsymbol{Q}(t_k) = \boldsymbol{q}, \boldsymbol{H}(t_k) = \boldsymbol{h}\right] \\ &\leqslant \sum_{\tau=D_k}^{D_k+D_{k+1}-1} \hat{E}_{t_k}\left[\sum_i 2Q_i(t_k+\tau)A_i(t_k+\tau)\right] \end{aligned} \tag{5}$$

$$- \sum_{\tau=D_k}^{D_k+D_{k+1}-1} \hat{E}_{t_k}\left[\sum_i 2Q_i(t_k+\tau)\sum_j \mathbb{1}_{i,j}(t_k+\tau)\right] \tag{6}$$

$$+ O(g(\gamma)),$$

where $\hat{E}_{t_k}$ is a shorthand for expectation conditioned on $\{\boldsymbol{Q}(t_k) = \boldsymbol{q}, \boldsymbol{H}(t_k) = \boldsymbol{h}\}$. In order to obtain a negative Lyapunov drift, we analyze the above two terms, the arrival term (5) and the service term (6). By using the inequality in (3), the arrival term (5) can be upper bounded by

$$(5) \leqslant 2\sum_j \sum_{\tau=D_k}^{D_k+D_{k+1}-1} \max_i q_i\mu_{i,j}(t_k+\tau) \tag{7}$$

$$+ O(g(\gamma)^2) - 2\delta \sum_{\tau=D_k}^{D_k+D_{k+1}-1} \hat{E}_{t_k}\left[\sum_i Q_i(t_k+\tau)\right], \tag{8}$$

where we hope that the term (7) can be later canceled out by the lower bound of the service term (6). In the next subsection, we analyze the service term (6).

5.2 Bounding the Service Term

Notice that the service term (6) is a sum over all servers j. Let us first fix one j and analyze the per-server service term:

$$\sum_{\tau=D_k}^{D_k+D_{k+1}-1} \hat{E}_{t_k}\left[\sum_i Q_i(t_k+\tau)\mathbb{1}_{i,j}(t_k+\tau)\right]. \tag{9}$$

Bounding the per-server service term (9) takes several steps.

Step 1: Concentration of Service Rates The first step is to prove a concentration result regarding the deviation of the estimates of the service rates $\hat{\mu}_{i,j}(t)$ from the true service rates $\mu_{i,j}(t)$. Consider a concentration event as follows:

$$\mathcal{E}_{t_k,j} := \Bigg\{\text{for all } i, \tau \in \left[D_k - \frac{g(\gamma)}{8}, D_k + D_{k+1} - 1\right], \\ |\hat{\mu}_{i,j}(t_k+\tau) - \mu_{i,j}(t_k+\tau)| \leqslant b_{i,j}(t_k+\tau)\Bigg\}. \tag{10}$$

Lemma 1. $\Pr(\mathcal{E}^c_{t_k,j} \,|\mathbf{Q}(t_k) = \mathbf{q}, \mathbf{H}(t_k) = \mathbf{h}) \leqslant \frac{I}{g(\gamma)^2}$ *for any* $k \geqslant 0$.

Lemma 1 shows that the deviation of the estimated service rates from the true service rates is bounded by the UCB bonus with high probability conditioned on the queue length $\boldsymbol{Q}(t_k)$ and $\boldsymbol{H}(t_k)$. Proving Lemma 1 is the most challenging part of our proof. Lemma 1 cannot be proved by simply using the Hoeffding inequality and the union bound like in the traditional analysis of UCB algorithms. There are three main difficulties. First, the probability is conditioned on the queue length in the previous interval, which is related to the service times before the previous interval. Thanks to the relation between the discount factor γ and the length $g(\gamma)$ of each interval, the contribution of the service times before the previous interval to the current estimate $\hat{\phi}_{i,j}(t)$ is negligible and can be bounded. Another difficulty is that $\hat{\phi}_{i,j}(t)$ is the discounted sum of previous service times and the summation is taken over the time slots in which there is job completion. Those time slots are random variables, which implies that the discount coefficients of those service times are also random. Also, $\hat{N}_{i,j}(t_k+\tau)$ is the sum of some discount coefficients, which is a random variable that takes values in the real line while in the standard MAB problem this is just a random integer. Therefore, taking union bound over $\hat{N}_{i,j}(t_k+\tau)$ like in the standard MAB analysis does not work in our setting. To deal with this difficulty, we first transform the summation into a summation over the time slots in which there is a job starting, and then use a Hoeffding-type inequality for self-normalized means with a random number of summands (Garivier and Moulines, 2008, Theorem 18)(Garivier and Moulines, 2011) to obtain a concentration bound. Another issue is that the mean service times are time-varying and the estimate of the mean service time in the current time slot is based on the actual service times in previous time slots. We utilize the first condition in Assumption 1 to solve this time-varying issue.

Using Lemma 1, we can show that (9) can be bounded by

$$(9) \geqslant -\frac{2I}{g(\gamma)^2} \sum_{\tau=D_k}^{D_k+D_{k+1}-1} \hat{E}_{t_k}\left[\sum_i Q_i(t_k+\tau)\right] - C \tag{11}$$

$$+ \hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1} \sum_i Q_i(t_k+\tau)\mathbb{1}_{i,j}(t_k+\tau)\,|\mathcal{E}_{t_k,j}\right], \tag{12}$$

where C is some constant. When $g(\gamma)$ is sufficiently large, (11) is negligible. (12) can be transformed into:

$$(12) = \hat{E}_{t_k}\Bigg[\sum_{\tau=D_k}^{D_k+D_{k+1}-1} Q_{I_j(t_k+\tau)}(t_k+\tau) \\ \mu_{I_j(t_k+\tau),j}(f_j(t_k+\tau))\frac{\mathbb{1}_{I_j(t_k+\tau),j}(t_k+\tau)}{\mu_{I_j(t_k+\tau),j}(f_j(t_k+\tau))}\,|\mathcal{E}_{t_k,j}\Bigg], \tag{13}$$

where $f_j(t)$ denotes the starting time of the job that is being served at server j in time slot t.

Step 2: Bound the Product of Queue Length and Service Rate We next bound the product of queue length and service rate, i.e., $Q_{I_j(t_k+\tau)}(t_k+\tau)\mu_{I_j(t_k+\tau),j}(f_j(t_k+\tau))$ in (13). Since the algorithm picks the largest product of queue length and UCB, this term can be lower bounded by $\max_i Q_i(t_k)\mu_{i,j}(f_j(t_k+\tau))$ minus some term containing

the UCB bonuses. Substituting this lower bound back to (13), we obtain

$$(12) \geqslant \hat{E}_{t_k}\Bigg[\sum_{\tau=D_k}^{D_k+D_{k+1}-1} \max_i q_i\mu_{i,j}(f_j(t_k+\tau)) \frac{\mathbb{1}_{I_j(t_k+\tau),j}(t_k+\tau)}{\mu_{I_j(t_k+\tau),j}(f_j(t_k+\tau))}\,|\mathcal{E}_{t_k,j}\Bigg] \quad (14)$$

$$- C\left(\sum_i q_i\right)\text{SumUCB} - O(g(\gamma)^2), \quad (15)$$

where SumUCB is the sum of UCB bonuses defined by

$$\text{SumUCB} := \sum_{\tau=D_k}^{D_k+D_{k+1}-1}\Bigg[b_{I_j(t_k+\tau),j}(f_j(t_k+\tau)) \eta_j(f_j(t_k+\tau))\mathbb{1}_{I_j(t_k+\tau),j}(t_k+\tau)\Bigg].$$

Lemma 2. *SumUCB* $\leqslant 99IU_{\text{S}}\sqrt{g(\gamma)}\log g(\gamma)$*, for any* j *and* $k \geqslant 0$.

The main difficulty in proving Lemma 2 is in obtaining a lower bound for $\hat{N}_{i,j}(t)$. Our method is to divide the interval into $\log g(\gamma)$ sub-intervals with length $\frac{g(\gamma)}{\log g(\gamma)}$ so that the discount coefficients can be lower bounded by a constant in each sub-interval. The bound in Lemma 2 is sublinear with respect to $g(\gamma)$, which is approximately equal to the length of the interval. Thanks to Lemma 2, the term (15) is negligible compared to the negative term in (8) if $g(\gamma)$ is sufficiently large.

Step 3: Bound the Weighted Sum of Job Completion Indicators The next step is to bound the weighted sum of job completion indicators (14), where $1/\mu_{I_j(t_k+\tau),j}(f_j(t_k+\tau))$ are mean service times. Intuitively, if we replace the mean service times with actual service times, the sum should not change too much. This concentration result can be proved using the same Hoeffding-type inequality for self-normalized means. The sum with actual service times is close to the sum of the weights over the time slots, i.e., $\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\max_i q_i\mu_{i,j}(t_k+\tau)$, if $\mu_{i,j}$ does not change too much within the duration of each service (the second condition in Assumption 1). That is, (14) $\gtrapprox$ $\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\max_i q_i\mu_{i,j}(t_k+\tau)$, where "$\gtrsim$" means that we drop some negligible terms.

Substituting the above bound into (14) and then back into (12), we have (9) $\gtrapprox$ $\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\max_i q_i\mu_{i,j}(t_k+\tau) - O(g(\gamma)^2)$. Summing over all servers j, we have the lower bound for the service term (6), i.e.,

$$(6) \gtrapprox 2\sum_j \sum_{\tau=D_k}^{D_k+D_{k+1}-1} \max_i q_i\mu_{i,j}(t_k+\tau) - O(g(\gamma)^2). \quad (16)$$

Substituting (16) into (6) and then substituting (7) and (8) into (5), we have

$$E_{t_k}\left[L(t_{k+1}+D_{k+1}) - L(t_{k+1})\right] \leqslant -\delta \sum_{\tau=D_k}^{D_k+D_{k+1}-1} E_{t_k}\left[\sum_i Q_i(t_k+\tau)\right] + O(g(\gamma)^2).$$

Finally, by doing a telescoping sum over all the intervals, we obtain the result in Theorem 1.

6 SIMULATION RESULTS

In this section, we evaluate the proposed algorithm numerically through simulation. We compare the proposed *MaxWeight with discounted UCB* algorithm with several baselines, including the *frame-based MaxWeight* algorithm (Stahlbuhk et al., 2019) and *DAM.UCB* algorithm (Freund et al., 2022). We also compare our algorithm with two MaxWeight algorithms using empirical mean (*MaxWeight with EM*) and discounted empirical mean (*MaxWeight with discounted EM*) as the estimated service rates.

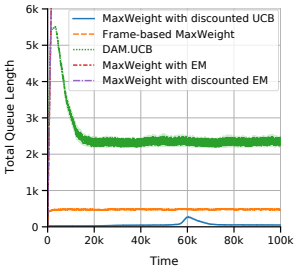
We consider a system with 10 job types and 10 servers. We compare the algorithms in the following four settings, stationary, nonstationary aperiodic, nonstationary periodic, and nonstationary periodic with a larger period. The simulation results are averaged over 100 runs. More details about the settings and parameters can be found in Appendix D. The results are shown in Fig. 3. We also present the same set of figures with a larger range of Y-axis in Appendix D.3, which show the missing parts of the curves.

Fig. 3a shows the results for the stationary setting, where the arrival rates and the service rates are time-invariant. As seen in the figure, the queue lengths of *MaxWeight with EM* and *MaxWeight with discounted EM* increase very fast and unstable and exceed 6000 after time slot 1520 and 1502, respectively, while *MaxWeight with discounted UCB*, *frame-based MaxWeight* and *DAM.UCB* are stable. The reason is that the empirical mean method lacks exploration so the system may "locks in" in a state with poor estimation and wrong scheduling decision. The queue length of *frame-based MaxWeight* is significantly larger than that of our algorithm because *frame-based MaxWeight* restarts the estimation and UCB of service rates at the beginning of every frame, which causes poor estimation. Another reason is that *frame-based MaxWeight* uses the queue length at the beginning of each frame to make decisions, which leads to wrong decisions in the frame because the queue length information becomes outdated. The queue length of *DAM.UCB* is several orders of magnitude larger than that of *MaxWeight with discounted UCB* because *DAM.UCB* uses the same schedule in each frame (called epoch in (Freund et al., 2022)), which also causes wrong decisions due to outdated information. We believe that the key reason why our algorithm performs

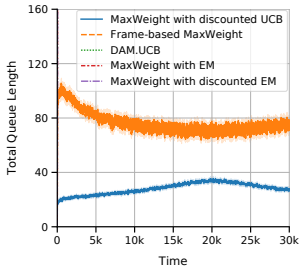
the best is that we continuously update both learning and scheduling decisions.

Fig. 3b shows the results for the nonstationary aperiodic setting and Fig. 3c and Fig. 3d show the results for the nonstationary periodic setting. Similar to the stationary setting, in all three cases, the queue lengths of *MaxWeight with EM*, *MaxWeight with discounted EM*, and *DAM.UCB* quickly exceed the Y-axis limits. *DAM.UCB* does not perform well because the service rates are changing over time but the algorithm is learning the service rates using outdated samples. While *frame-based MaxWeight* is stable, its queue length is larger and the oscillation is wilder. Note that the period of the setting of Fig. 3d is 10 times as large as that of Fig. 3c. As seen in the figures, for *frame-based MaxWeight*, both the amplitude of the oscillation and the peak value become larger when the period becomes larger, while the amplitude of the oscillation and the peak value for our algorithm remain approximately the same and even smaller. The reason is that our algorithm can quickly adapt to the changing statistics thanks to the discount factor.

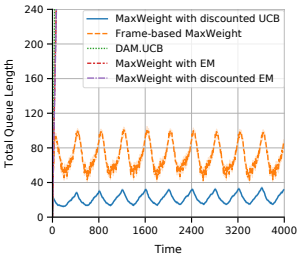
We also did some simulations of *MaxWeight with discounted UCB* algorithm with different $g(\gamma)$, which can be found in Fig. 5 in Appendix D. The results show that the proposed algorithm is robust to the value of γ.



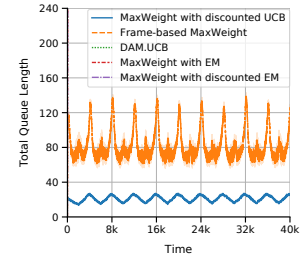
(a) Stationary. The queue lengths of *MaxWeight with EM* and *MaxWeight with discounted EM* exceed 6k after time slot 1520 and 1502, respectively.



(b) Nonstationary aperiodic. The queue lengths of *MaxWeight with EM*, *MaxWeight with discounted EM*, and *DAM.UCB* exceed 160 after time slot 42, 44, and 26, respectively.



(c) Nonstationary periodic: period=400 for arrival rates, period=800 for service rates. The queue lengths of *MaxWeight with EM*, *MaxWeight with discounted EM*, and *DAM.UCB* exceed 240 after time slot 65, 69, and 38, respectively.



(d) Nonstationary periodic: period=4k for arrival rates, period=8k for service rates. The queue lengths of *MaxWeight with EM*, *MaxWeight with discounted EM*, and *DAM.UCB* exceed 240 after time slot 61, 61, and 37, respectively.

Figure 3: Simulation Results.

7 CONCLUSIONS

This paper considered scheduling in multi-server queueing systems with unknown arrival and service statistics, and proposed a new scheduling algorithm, MaxWeight with discounted UCB. Based on the Lyapunov drift analysis and concentration inequalities of self-normalized means, we proved that MaxWeight with discounted UCB guarantees queue stability (in the mean) when the arrival rates are strictly within the service capacity region. This result holds both for stationary systems and nonstationary systems.

Acknowledgements

The work of Zixian Yang and Lei Ying is supported in part by NSF under grants 2001687, 2112471, 2207548, and 2228974. The work of R. Srikant is supported in part by NSF CCF 22-07547, NSF CNS 21-06801, NSF CCF 1934986, ONR N00014-19-1-2566, and ARO W911NF-19-1-0379.

References

Andrews, M., Jung, K., and Stolyar, A. (2007). Stability of the max-weight routing and scheduling protocol in dynamic networks and at critical loads. In *Proc. Ann. ACM Symp. Theory of Computing (STOC)*, pages 145–154.

Choudhury, T., Joshi, G., Wang, W., and Shakkottai, S. (2021). Job dispatching policies for queueing systems with unknown service rates. In *Proc. ACM Int. Symp. Mobile Ad Hoc Networking and Computing (MobiHoc)*, pages 181–190.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A probabilistic theory of pattern recognition*. Springer-Verlag.

Eryilmaz, A. and Srikant, R. (2012). Asymptotically tight steady-state queue length bounds implied by drift conditions. *Queueing Syst.*, 72(3-4):311–359.

Freund, D., Lykouris, T., and Weng, W. (2022). Efficient decentralized multi-agent learning in asymmetric queuing systems. In *Proc. Conf. Learning Theory (COLT)*, volume 178, pages 4080–4084.

Garivier, A. and Moulines, E. (2008). On upper-confidence bound policies for non-stationary bandit problems. *arXiv preprint arXiv:0805.3415*.

Garivier, A. and Moulines, E. (2011). On upper-confidence bound policies for switching bandit problems. In *Int. Conf. Algorithmic Learning Theory (ALT)*, pages 174–188. Springer.

Hsu, W.-K., Xu, J., Lin, X., and Bell, M. R. (2022). Integrated online learning and adaptive control in queueing

systems with uncertain payoffs. *Operations Research*, 70(2):1166–1181.

Kang, W. and Williams, R. (2013). Diffusion approximation for an input-queued switch operating under a maximum weight matching policy. *Stoch. Syst.*, 2(2):277–321.

Kocsis, L. and Szepesvári, C. (2006). Discounted UCB. In *2nd PASCAL Challenges Workshop*, volume 2.

Krishnasamy, S., Akhil, P. T., Arapostathis, A., Sundaresan, R., and Shakkottai, S. (2018a). Augmenting max-weight with explicit learning for wireless scheduling with switching costs. *IEEE/ACM Trans. Netw.*, 26(6):2501–2514.

Krishnasamy, S., Arapostathis, A., Johari, R., and Shakkottai, S. (2018b). On learning the $c\mu$ rule in single and parallel server networks. In *Proc. Annu. Allerton Conf. Communication, Control and Computing*.

Krishnasamy, S., Sen, R., Johari, R., and Shakkottai, S. (2021). Learning unknown service rates in queues: A multiarmed bandit approach. *Operations Research*, 69(1):315–330. The conference version appeared in NeurIPS 2016.

Li, F., Liu, J., and Ji, B. (2019). Combinatorial sleeping bandits with fairness constraints. In *Proc. IEEE Int. Conf. Computer Communications (INFOCOM)*, pages 1702–1710.

Liu, X., Li, B., Shi, P., and Ying, L. (2021). An efficient pessimistic-optimistic algorithm for stochastic linear bandits with general constraints. In *Advances Neural Information Processing Systems (NeurIPS)*.

Maguluri, S. T. and Srikant, R. (2016). Heavy traffic queue length behavior in a switch under the maxweight algorithm. *Stoch. Syst.*, 6(1):211–250.

Neely, M. J., Rager, S. T., and La Porta, T. F. (2012). Max weight learning algorithms for scheduling in unknown environments. *IEEE Trans. Autom. Control*, 57(5):1179–1191.

Shah, D. and Wischik, D. (2007). Heavy traffic analysis of optimal scheduling algorithms for switched networks. Submitted to Annals of Applied Probability.

Srikant, R. and Ying, L. (2014). *Communication Networks: An Optimization, Control and Stochastic Networks Perspective*. Cambridge University Press.

Stahlbuhk, T., Shrader, B., and Modiano, E. (2019). Learning algorithms for scheduling in wireless networks with unknown channel statistics. *Ad Hoc Networks*, 85:131–144.

Stolyar, A. L. (2004). MaxWeight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *Adv. in Appl. Probab.*, 14(1).

Tassiulas, L. and Ephremides, A. (1992). Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Trans. Autom. Control*, 37:1936–1948.

Tassiulas, L. and Ephremides, A. (1993). Dynamic server allocation to parallel queues with randomly varying connectivity. *IEEE Trans. Inf. Theory*, 39:466–478.

Yekkehkhany, A. and Nagi, R. (2020). Blind gb-pandas: A blind throughput-optimal load balancing algorithm for affinity scheduling. *IEEE/ACM Transactions on Networking*, 28(3):1199–1212.

TABLE OF CONTENTS IN THE SUPPLEMENTARY MATERIALS

In the supplementary materials, we provide a counter-example of MaxWeight with empirical mean algorithm, the complete proof of Theorem 1, the proofs of all the lemmas, and additional details of the simulations. The contents are listed as follows:

A A COUNTER-EXAMPLE OF MAXWEIGHT WITH EMPIRICAL MEAN ALGORITHM

In this section, we will present an example showing that the MaxWeight with empirical mean algorithm is unstable.

Consider a multi-server system with two servers and two job types with the following statistics:

$$\Pr\left(S_{i,j}=1\right)=0.99, \quad \Pr\left(S_{i,j}=100\right)=0.01 \quad \text{for} \quad i=j$$

and

$$\Pr\left(S_{i,j}=10\right)=1 \quad \text{for} \quad i \neq j,$$

where $S_{i,j}$ is the service time of type i jobs at server j. We further assume the following job arrival process: $A_i(t)=1$ for any i and any $t=1,3,\ldots$ and $A_i(t)=0$ for any i and $t=2,4,\ldots$. We next consider the queue lengths over time under MaxWeight with empirical mean, and assume the algorithm uses $\hat{\mu}_{i,j}=1$ as a default value if there is no data sample for $S_{i,j}$.

- Time slot 1: A type-i job is scheduled at server i and $S_{i,i}=100$ for $i=1,2$ which occurs with probability 0.01.
- Time slot 101: Both queues have 49 jobs. We have estimated $\hat{\mu}_{i,i}=0.01$ and $\hat{\mu}_{i,j}=1$ $(i \neq j)$ as the default value. The algorithm now schedules a type-i job to server j for $j \neq i$.
- Time slot 111 : Both queues have 53 jobs. The estimated service rates are $\hat{\mu}_{i,i}=0.01$ and $\hat{\mu}_{i,j}=0.1$ for $i \neq j$. Based on MaxWeight with mean-service-rate, the scheduler schedules type-i jobs to server j for $i \neq j$.
- Time slot >111 : Since $S_{i,j}$ is a constant for $i \neq j$, the estimated service rates do not change after the jobs are completed. Since the estimated service rates do not change as long as type-i jobs are scheduled on server j such that $i \neq j$, the schedule decisions also remain the same such that type-i jobs are continuously scheduled to server j for $i \neq j$. Since it takes 10 time slots to finish a job and there is a job arrival every two slots, both queues go to infinity.

Note that if we schedule type-i jobs to server i, the mean queue lengths are bounded because in this case, the mean service time is 1.99 time slots and the arrival rate is one job every two time slots. □

From the example above, we can see that the problem of using empirical mean is that the initial bad samples led to a poor estimation of $\mu_{i,j}$, which led to poor scheduling decisions. Since the scheduler only gets new samples from the served jobs, it was not able to correct the wrong estimate of $\hat{\mu}_{i,i}=0.01$ when type-i jobs are no long routed to server i after time slot 101. Therefore, the system was "locked in" in a state with poor estimation and wrong scheduling decisions, which led to instability.

B PROOF OF THEOREM 1

In this section, we present the complete proof of Theorem 1. Fig. 4 shows the proof raodmap of Theorem 1. Before presenting the proof, we define a few additional notations. In the proof, if server j is not available at the beginning of time slot t, i.e., $\sum_i M_{i,j}(t)>0$, we let $\hat{i}_j^*(t)=0$. Let $T:=g(\gamma)$ for ease of notation.

We now present the proof of Theorem 1 in the following subsections.

B.1 Dividing the Time Horizon

Firstly, we want to divide the time horizon into intervals. We assume $\frac{T}{8}$ is an integer without loss of generality. Since $\boldsymbol{\lambda}+\delta\mathbf{1} \in \mathcal{C}(W)$, for any time slot τ, there exists a $w(\tau)$ that satisfies the inequality in the capacity region definition (3). Let $\tau_0(t):=t$, $\tau_l(t):=\tau_{l-1}(t)+w(\tau_{l-1}(t))$ for $l \geqslant 1$. Define $D(t)$ such that

$$D(t)=\min_n \sum_{l=0}^{n} w(\tau_l(t)) \qquad \text{s.t.} \ \sum_{l=0}^{n} w(\tau_l(t)) \geqslant \frac{T}{2}.$$

Denote by $n^*(t)$ the optimal solution to the above optimization problem. Note that $n^*(t)$ and $D(t)$ are fixed numbers rather than random variables for a given t. We have the following upper and lower bounds for $D(t)$:

Lemma 3. *Suppose $W \leqslant \frac{T}{2}$. Then $\frac{T}{2} \leqslant D(t) \leqslant \frac{T}{2}+W \leqslant T$ for any t.*

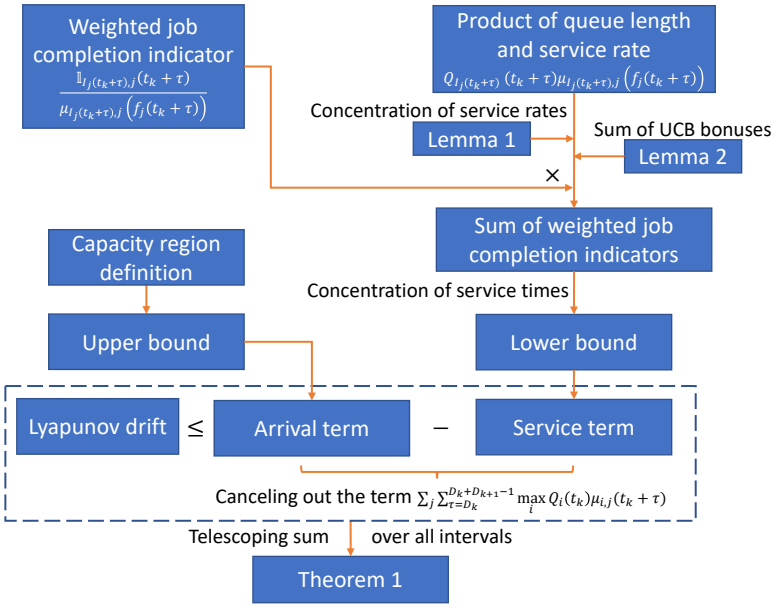


Figure 4: The proof roadmap of Theorem 1.

Proof of this lemma can be found in Section C.3. Let $t_0 = 0$ and $t_k = t_{k-1} + D(t_{k-1})$ for $k \geqslant 1$. Let $D_k := D(t_k)$ for simplicity. Then the time horizon can be divided into intervals with length $D_0, D_1, \ldots, D_k, \ldots$, where the k^{th} interval is $[t_k, t_{k+1}]$. We remark that this partition of the time horizon into time intervals is for the analysis only. The proposed algorithm does not need to know this partition and does not use the time interval information for scheduling and learning.

In the next subsection, we will analyze and decompose the Lyapunov Drift in each interval.

B.2 Decomposing the Lyapunov Drift

Consider the Lyapunov function $L(t) := \sum_i Q_i^2(t)$. We first consider the Lyapunov drift for the interval $[t_{k+1}, t_{k+1} + D_{k+1}]$ given the queue length $\boldsymbol{Q}(t_k)$ and $\boldsymbol{H}(t_k)$. We analyze the drift conditioned on $\boldsymbol{Q}(t_k)$ and $\boldsymbol{H}(t_k)$ instead of $\boldsymbol{Q}(t_{k+1})$ and $\boldsymbol{H}(t_{k+1})$ to weaken the dependence of the UCB bonuses and the estimated service rates on the conditional values. We have

$$\begin{aligned}
&E\left[L(t_{k+1} + D_{k+1}) - L(t_{k+1}) \,|\boldsymbol{Q}(t_k) = \boldsymbol{q}, \boldsymbol{H}(t_k) = \boldsymbol{h}\right] \\
=&E\left[L(t_k + D_k + D_{k+1}) - L(t_k + D_k) \,|\boldsymbol{Q}(t_k) = \boldsymbol{q}, \boldsymbol{H}(t_k) = \boldsymbol{h}\right] \\
=&\sum_{\tau=D_k}^{D_k+D_{k+1}-1} E\left[L(t_k + \tau + 1) - L(t_k + \tau) \,|\boldsymbol{Q}(t_k) = \boldsymbol{q}, \boldsymbol{H}(t_k) = \boldsymbol{h}\right]. \qquad (17)
\end{aligned}$$

We first look at each term in the summation above. Note that by the queue dynamic (1) we can obtain the following upper bound for $Q_i(t+1)$:

Lemma 4. *For any i, t, $Q_i(t+1) \leqslant \max\left\{J, Q_i(t) + A_i(t) - \sum_j \mathbb{1}_{i,j}(t)\right\}$.*

Proof of this lemma can be found in Section C.4. Denote by $\hat{E}_{t_k}[\cdot]$ the conditional expectation $E[\cdot\,|\boldsymbol{Q}(t_k) = \boldsymbol{q}, \boldsymbol{H}(t_k) = \boldsymbol{h}]$. By Lemma 4, we have

$$\begin{aligned}
&\hat{E}_{t_k}\left[L(t_k + \tau + 1) - L(t_k + \tau)\right] = \hat{E}_{t_k}\left[\sum_i \left(Q_i^2(t_k + \tau + 1) - Q_i^2(t_k + \tau)\right)\right] \\
\leqslant&\hat{E}_{t_k}\left[\sum_i\left[\max\left\{J^2, \left(Q_i(t_k + \tau) + A_i(t_k + \tau) - \sum_j \mathbb{1}_{i,j}(t_k + \tau)\right)^2\right\} - Q_i^2(t_k + \tau)\right]\right] \\
\leqslant&\hat{E}_{t_k}\left[\sum_i\left[J^2 + \left(Q_i(t_k + \tau) + A_i(t_k + \tau) - \sum_j \mathbb{1}_{i,j}(t_k + \tau)\right)^2 - Q_i^2(t_k + \tau)\right]\right]
\end{aligned}$$

$$=\hat{E}_{t_k}\left[\sum_i 2Q_i(t_k+\tau)\big(A_i(t_k+\tau)-\sum_j \mathbb{1}_{i,j}(t_k+\tau)\big)\right]+\hat{E}_{t_k}\left[\sum_i\big(A_i(t_k+\tau)-\sum_j \mathbb{1}_{i,j}(t_k+\tau)\big)^2\right]+IJ^2 \tag{18}$$

where the second inequality is due to the fact that $\max\{x^2,y^2\}\leqslant x^2+y^2$, and the second term in the last line can be bounded as follows:

$$\begin{aligned}&\sum_i\big(A_i(t_k+\tau)-\sum_j \mathbb{1}_{i,j}(t_k+\tau)\big)^2\leqslant\sum_i\big(\max\big\{A_i(t_k+\tau),\sum_j \mathbb{1}_{i,j}(t_k+\tau)\big\}\big)^2\\ \leqslant&\sum_i(A_i(t_k+\tau))^2+\sum_i\big(\sum_j \mathbb{1}_{i,j}(t_k+\tau)\big)^2\leqslant IU_{\mathrm{A}}^2+\big[\sum_i\sum_j \mathbb{1}_{i,j}(t_k+\tau)\big]^2\leqslant IU_{\mathrm{A}}^2+J^2,\end{aligned}$$

where the last two steps are due to the fact that $A_i(t_k+\tau)\leqslant U_{\mathrm{A}}$ and $\sum_i\sum_j \mathbb{1}_{i,j}(t_k+\tau)\leqslant J$. Hence, from (18), we have

$$\begin{aligned}&\hat{E}_{t_k}\left[L(t_k+\tau+1)-L(t_k+\tau)\right]\\ \leqslant&\hat{E}_{t_k}\left[\sum_i 2Q_i(t_k+\tau)A_i(t_k+\tau)\right]-\hat{E}_{t_k}\left[\sum_i 2Q_i(t_k+\tau)\sum_j \mathbb{1}_{i,j}(t_k+\tau)\right]+IU_{\mathrm{A}}^2+J^2+IJ^2.\end{aligned}$$

Substituting the above inequality into (17), we have

$$\begin{aligned}&E\left[L(t_k+D_k+D_{k+1})-L(t_k+D_k)\,|\boldsymbol{Q}(t_k)=\boldsymbol{q},\boldsymbol{H}(t_k)=\boldsymbol{h}\right]\\ \leqslant&\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\hat{E}_{t_k}\left[\sum_i 2Q_i(t_k+\tau)A_i(t_k+\tau)\right] &(19)\\ &-\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\hat{E}_{t_k}\left[\sum_i 2Q_i(t_k+\tau)\sum_j \mathbb{1}_{i,j}(t_k+\tau)\right]+(IU_{\mathrm{A}}^2+J^2+IJ^2)T &(20)\end{aligned}$$

where the inequality uses the the upper bound on D_{k+1} in Lemma 3. We will next find the bounds for the arrival term (19) and the service term (20).

In the next subsection, we will bound the arrival term (19).

B.3 Bounding the Arrival Term

We first analyze the arrival term (19). We have

$$\begin{aligned}(19)=&\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\hat{E}_{t_k}\left[E\left[\sum_i 2Q_i(t_k+\tau)A_i(t_k+\tau)\,|\boldsymbol{Q}(t_k+\tau),\boldsymbol{Q}(t_k)=\boldsymbol{q},\boldsymbol{H}(t_k)=\boldsymbol{h}\right]\right]\\ =&\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\hat{E}_{t_k}\left[\sum_i 2Q_i(t_k+\tau)E\left[A_i(t_k+\tau)\,|\boldsymbol{Q}(t_k+\tau),\boldsymbol{Q}(t_k)=\boldsymbol{q},\boldsymbol{H}(t_k)=\boldsymbol{h}\right]\right]\\ =&\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\hat{E}_{t_k}\left[\sum_i 2Q_i(t_k+\tau)\lambda_i(t_k+\tau)\right],\end{aligned}$$

where the first equality is by law of iterated expectation and the last equality is due to the fact that $A_i(t_k+\tau)$ is independent of $\boldsymbol{Q}(t_k+\tau)$, $\boldsymbol{Q}(t_k)$, and $\boldsymbol{H}(t_k)$. By adding and subtracting δ, we have

$$(19)=\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\hat{E}_{t_k}\left[\sum_i 2Q_i(t_k+\tau)(\lambda_i(t_k+\tau)+\delta)\right]-2\delta\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\hat{E}_{t_k}\left[\sum_i Q_i(t_k+\tau)\right]. \tag{21}$$

By the queue dynamics (1) and the bounds on the arrival rate and service rate, we have the following bounds on the difference between queue lengths in two different time slots:

Lemma 5. *For any $t, i, \tau\geqslant 0$, we have*

1. $Q_i(t)-J\tau\leqslant Q_i(t+\tau)\leqslant Q_i(t)+\tau U_{\mathrm{A}}$;

2. $\sum_i Q_i(t+\tau) \geqslant \sum_i Q_i(t) - J\tau$.

Proof of this lemma can be found in Section C.5. By Lemma 5, we have

$$Q_i(t_k+\tau) \leqslant Q_i(t_k) + \tau U_{\mathrm{A}} \leqslant Q_i(t_k) + 2TU_{\mathrm{A}} \tag{22}$$

where the last inequality holds since $\tau \leqslant D_k + D_{k+1} - 1 \leqslant 2T$ by Lemma 3. Then, substituting (22) into (21), we have

$$\begin{aligned}(19) \leqslant & \sum_{\tau=D_k}^{D_k+D_{k+1}-1} \sum_i 2q_i(\lambda_i(t_k+\tau)+\delta) + \sum_{\tau=D_k}^{D_k+D_{k+1}-1} \sum_i 4TU_{\mathrm{A}}(\lambda_i(t_k+\tau)+\delta) \\ & - 2\delta \sum_{\tau=D_k}^{D_k+D_{k+1}-1} \hat{E}_{t_k}\left[\sum_i Q_i(t_k+\tau)\right].\end{aligned} \tag{23}$$

Since $\boldsymbol{\lambda}+\delta\mathbf{1} \in \mathcal{C}(W)$, by the definitions of t_{k+1} and D_{k+1}, we have

$$\begin{aligned}&\sum_{\tau=D_k}^{D_k+D_{k+1}-1} (\lambda_i(t_k+\tau)+\delta) = \sum_{\tau=t_k+D_k}^{t_k+D_k+D_{k+1}-1} (\lambda_i(\tau)+\delta) = \sum_{\tau=t_{k+1}}^{t_{k+1}+D_{k+1}-1} (\lambda_i(\tau)+\delta) \\ =& \sum_{\tau=t_{k+1}}^{t_{k+1}+\sum_{l=0}^{n^*(t_{k+1})} w(\tau_l(t_{k+1}))-1} (\lambda_i(\tau)+\delta) = \sum_{l=0}^{n^*(t_{k+1})} \sum_{\tau=t_{k+1}+\sum_{l'=0}^{l-1} w(\tau_{l'}(t_{k+1}))}^{t_{k+1}+\sum_{l'=0}^{l} w(\tau_{l'}(t_{k+1}))-1} (\lambda_i(\tau)+\delta).\end{aligned}$$

Then by the definitions of $\tau_l(t_{k+1})$ and $\mathcal{C}(W)$, we can bound the above term as follows:

$$\begin{aligned}&\sum_{\tau=D_k}^{D_k+D_{k+1}-1} (\lambda_i(t_k+\tau)+\delta) = \sum_{l=0}^{n^*(t_{k+1})} \sum_{\tau=\tau_0(t_{k+1})+\sum_{l'=0}^{l-1} w(\tau_{l'}(t_{k+1}))}^{\tau_0(t_{k+1})+\sum_{l'=0}^{l} w(\tau_{l'}(t_{k+1}))-1} (\lambda_i(\tau)+\delta) \\ =& \sum_{l=0}^{n^*(t_{k+1})} \sum_{\tau=\tau_l(t_{k+1})}^{\tau_l(t_{k+1})+w(\tau_l(t_{k+1}))-1} (\lambda_i(\tau)+\delta) \leqslant \sum_{l=0}^{n^*(t_{k+1})} \sum_{\tau=\tau_l(t_{k+1})}^{\tau_l(t_{k+1})+w(\tau_l(t_{k+1}))-1} \sum_j \alpha_{i,j}(\tau)\mu_{i,j}(\tau).\end{aligned}$$

In the same way, we can transform the double summations back to a single summation to obtain:

$$\sum_{\tau=D_k}^{D_k+D_{k+1}-1} (\lambda_i(t_k+\tau)+\delta) \leqslant \sum_{\tau=D_k}^{D_k+D_{k+1}-1} \sum_j \alpha_{i,j}(t_k+\tau)\mu_{i,j}(t_k+\tau).$$

Substituting the above bound back into (23), we have

$$\begin{aligned}(19) \leqslant & \sum_{\tau=D_k}^{D_k+D_{k+1}-1} \sum_i 2q_i \sum_j \alpha_{i,j}(t_k+\tau)\mu_{i,j}(t_k+\tau) + 4TU_{\mathrm{A}} \sum_{\tau=D_k}^{D_k+D_{k+1}-1} \sum_i \sum_j \alpha_{i,j}(t_k+\tau)\mu_{i,j}(t_k+\tau) \\ & - 2\delta \sum_{\tau=D_k}^{D_k+D_{k+1}-1} \hat{E}_{t_k}\left[\sum_i Q_i(t_k+\tau)\right].\end{aligned} \tag{24}$$

Note that

$$\sum_{\tau=D_k}^{D_k+D_{k+1}-1} \sum_i \sum_j \alpha_{i,j}(t_k+\tau)\mu_{i,j}(t_k+\tau) \leqslant \sum_{\tau=D_k}^{D_k+D_{k+1}-1} \sum_j \max_i \mu_{i,j}(t_k+\tau) \leqslant JD_{k+1} \leqslant JT, \tag{25}$$

where the first inequality is by $\sum_i \alpha_{i,j}(t_k+\tau) \leqslant 1$, the second inequality is by $\mu_{i,j}(t_k+\tau) \leqslant 1$, and the last inequality is by Lemma 3. Note that

$$\sum_i q_i \alpha_{i,j}(t_k+\tau)\mu_{i,j}(t_k+\tau) \leqslant \max_i q_i\mu_{i,j}(t_k+\tau) \sum_i \alpha_{i,j}(t_k+\tau) \leqslant \max_i q_i\mu_{i,j}(t_k+\tau), \tag{26}$$

where the last inequality is by $\sum_i \alpha_{i,j}(t_k+\tau) \leqslant 1$. Substituting (25) and (26) into (24), we have

$$(19) \leqslant 2\sum_j \sum_{\tau=D_k}^{D_k+D_{k+1}-1} \max_i q_i\mu_{i,j}(t_k+\tau) + 4T^2JU_{\mathrm{A}} - 2\delta \sum_{\tau=D_k}^{D_k+D_{k+1}-1} \hat{E}_{t_k}\left[\sum_i Q_i(t_k+\tau)\right]. \tag{27}$$

In the next subsection, we will bound the the service term (20).

B.4 Bounding the Service Term

Now we analyze the service term (20). Let us first fix a server j. We want to lower bound the following per-server service term:

$$\sum_{\tau=D_k}^{D_k+D_{k+1}-1} \hat{E}_{t_k}\left[\sum_i Q_i(t_k+\tau)\mathbb{1}_{i,j}(t_k+\tau)\right].$$

The process takes several steps, which are shown in the following.

B.4.1 Step 1: Adding the Concentration to the Condition

Denote by $\hat{P}_{t_k}(\cdot)$ the conditional probability $\Pr(\cdot|\mathbf{Q}(t_k)=\mathbf{q},\mathbf{H}(t_k)=\mathbf{h})$. Denote by $\hat{E}_{t_k}[\cdot|\mathcal{E}_{t_k,j}]$ the conditional expectation $E[\cdot|\mathbf{Q}(t_k)=\mathbf{q},\mathbf{H}(t_k)=\mathbf{h},\mathcal{E}_{t_k,j}]$. Then by Lemma 1, we have

$$\begin{aligned}
&\sum_{\tau=D_k}^{D_k+D_{k+1}-1} \hat{E}_{t_k}\left[\sum_i Q_i(t_k+\tau)\mathbb{1}_{i,j}(t_k+\tau)\right]\\
\geqslant&\hat{P}_{t_k}(\mathcal{E}_{t_k,j})\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\sum_i Q_i(t_k+\tau)\mathbb{1}_{i,j}(t_k+\tau)\,|\mathcal{E}_{t_k,j}\right]\\
\geqslant&\left(1-\frac{I}{T^2}\right)\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\sum_i Q_i(t_k+\tau)\mathbb{1}_{i,j}(t_k+\tau)\,|\mathcal{E}_{t_k,j}\right].
\end{aligned}\tag{28}$$

Using the bound (22) and the bound on D_{k+1} in Lemma 3, we have

$$\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\sum_i Q_i(t_k+\tau)\mathbb{1}_{i,j}(t_k+\tau)\,|\mathcal{E}_{t_k,j}\right]\leqslant T\sum_i q_i+2U_{\mathrm{A}}T^2I.\tag{29}$$

By Lemma 5 and Lemma 3, we can obtain the following bound on $\sum_i q_i$:

Lemma 6. $\sum_i q_i \leqslant \frac{1}{D_{k+1}}\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\hat{E}_{t_k}\left[\sum_i Q_i(t_k+\tau)+2JT\right]$.

Proof of this lemma can be found in Section C.6. By (29), Lemma 6, and Lemma 3, we have

$$\begin{aligned}
&\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\sum_i Q_i(t_k+\tau)\mathbb{1}_{i,j}(t_k+\tau)\,|\mathcal{E}_{t_k,j}\right]\\
\leqslant&2\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\hat{E}_{t_k}\left[\sum_i Q_i(t_k+\tau)\right]+2JT^2+2U_{\mathrm{A}}T^2I,
\end{aligned}\tag{30}$$

Substituting (30) into (28), we have

$$\begin{aligned}
&\sum_{\tau=D_k}^{D_k+D_{k+1}-1} \hat{E}_{t_k}\left[\sum_i Q_i(t_k+\tau)\mathbb{1}_{i,j}(t_k+\tau)\right]\\
\geqslant&\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\sum_i Q_i(t_k+\tau)\mathbb{1}_{i,j}(t_k+\tau)\,|\mathcal{E}_{t_k,j}\right]\\
&-\frac{2I}{T^2}\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\hat{E}_{t_k}\left[\sum_i Q_i(t_k+\tau)\right]-2IJ-2U_{\mathrm{A}}I^2.
\end{aligned}\tag{31}$$

Next, we want to lower bound the term $\hat{E}_{t_k}[\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\sum_i Q_i(t_k+\tau)\mathbb{1}_{i,j}(t_k+\tau)|\mathcal{E}_{t_k,j}]$ in (31) using $\boldsymbol{q}$ and $\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\hat{E}_{t_k}\left[\sum_i Q_i(t_k+\tau)\right]$. Note that $\mathbb{1}_{i,j}(t_k+\tau)=1$ can only happen for the queue to which server j is scheduled in time slot $t_k+\tau$, i.e., the queue $I_j(t_k+\tau)$. Hence, we have

$$\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\sum_i Q_i(t_k+\tau)\mathbb{1}_{i,j}(t_k+\tau)\,|\mathcal{E}_{t_k,j}\right]$$

$$=\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1} Q_{I_j(t_k+\tau)}(t_k+\tau)\mathbb{1}_{I_j(t_k+\tau),j}(t_k+\tau)\,|\mathcal{E}_{t_k,j}\right].$$

Define a mapping f_j that maps a time slot to another time slot such that if $y = f_j(x)$ then y is the time slot when server j picked the job that was being served at server j in time slot x. If server j was idling in time slot x, then let $f_j(x) = x$. That is, $f_j(x) \coloneqq \max\{t : t \leqslant x, \hat{i}_j^*(t) = I_j(x)\}$. Then

$$\begin{aligned}
&\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\sum_i Q_i(t_k+\tau)\mathbb{1}_{i,j}(t_k+\tau)\,|\mathcal{E}_{t_k,j}\right]\\
=&\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1} Q_{I_j(t_k+\tau)}(t_k+\tau)\mu_{I_j(t_k+\tau),j}(f_j(t_k+\tau))\frac{\mathbb{1}_{I_j(t_k+\tau),j}(t_k+\tau)}{\mu_{I_j(t_k+\tau),j}(f_j(t_k+\tau))}\,|\mathcal{E}_{t_k,j}\right].
\end{aligned} \tag{32}$$

B.4.2 Step 2: Bounding the Product of Queue Length and Service Rate

We next want to lower bound the term $Q_{I_j(t_k+\tau)}(t_k+\tau)\mu_{I_j(t_k+\tau),j}(f_j(t_k+\tau))$ in (32). The following analysis is conditioned on the concentration event $\mathcal{E}_{t_k,j}$. Since $U_{\mathrm{S}} \leqslant \frac{T}{8}$, we have $f_j(t_k+\tau) \geqslant t_k+\tau-U_{\mathrm{S}} \geqslant t_k+D_k-U_{\mathrm{S}} \geqslant t_k+D_k-\frac{T}{8}$. Also note that $f_j(t_k+\tau) \leqslant t_k+\tau \leqslant t_k+D_k+D_{k+1}-1$. Hence, we have

$$f_j(t_k+\tau) \in \left[t_k+D_k-\frac{T}{8}, t_k+D_k+D_{k+1}-1\right] \subseteq [t_k, t_k+2T-1], \tag{33}$$

where the inclusion is by Lemma 3. Then, by (33) and the definition of the concentration event $\mathcal{E}_{t_k,j}$ in (10), we have

$$\begin{aligned}
&Q_{I_j(t_k+\tau)}(t_k+\tau)\mu_{I_j(t_k+\tau),j}(f_j(t_k+\tau))\\
\geqslant&Q_{I_j(t_k+\tau)}(t_k+\tau)\left(\hat{\mu}_{I_j(t_k+\tau),j}(f_j(t_k+\tau))-b_{I_j(t_k+\tau),j}(f_j(t_k+\tau))\right)\\
=&Q_{I_j(t_k+\tau)}(t_k+\tau)\left(\hat{\mu}_{I_j(t_k+\tau),j}(f_j(t_k+\tau))+b_{I_j(t_k+\tau),j}(f_j(t_k+\tau))\right)\\
&-2Q_{I_j(t_k+\tau)}(t_k+\tau)b_{I_j(t_k+\tau),j}(f_j(t_k+\tau)).
\end{aligned}$$

Note that by Lemma 5 and the fact that $t_k+\tau-f_j(t_k+\tau) \leqslant U_{\mathrm{S}} \leqslant \frac{T}{8}$, we have $Q_{I_j(t_k+\tau)}(t_k+\tau) \geqslant Q_{I_j(t_k+\tau)}(f_j(t_k+\tau)) - \frac{JT}{8}$. Hence, we have

$$\begin{aligned}
&Q_{I_j(t_k+\tau)}(t_k+\tau)\mu_{I_j(t_k+\tau),j}(f_j(t_k+\tau))\\
\geqslant&Q_{I_j(t_k+\tau)}(f_j(t_k+\tau))\left(\hat{\mu}_{I_j(t_k+\tau),j}(f_j(t_k+\tau))+b_{I_j(t_k+\tau),j}(f_j(t_k+\tau))\right)-\frac{JT}{4}\\
&-2Q_{I_j(t_k+\tau)}(t_k+\tau)b_{I_j(t_k+\tau),j}(f_j(t_k+\tau))
\end{aligned} \tag{34}$$

since $\hat{\mu}_{I_j(t_k+\tau),j}(f_j(t_k+\tau)) + b_{I_j(t_k+\tau),j}(f_j(t_k+\tau)) \leqslant 2$. By Line 18 in Algorithm 1, we have

$$\begin{aligned}
&Q_{I_j(t_k+\tau)}(f_j(t_k+\tau))\left(\hat{\mu}_{I_j(t_k+\tau),j}(f_j(t_k+\tau))+b_{I_j(t_k+\tau),j}(f_j(t_k+\tau))\right)\\
=&\max_i Q_i(f_j(t_k+\tau))\left(\hat{\mu}_{i,j}(f_j(t_k+\tau))+b_{i,j}(f_j(t_k+\tau))\right).
\end{aligned} \tag{35}$$

Combining (34) and (35), we have

$$\begin{aligned}
&Q_{I_j(t_k+\tau)}(t_k+\tau)\mu_{I_j(t_k+\tau),j}(f_j(t_k+\tau))\\
\geqslant&\max_i Q_i(f_j(t_k+\tau))\left(\hat{\mu}_{i,j}(f_j(t_k+\tau))+b_{i,j}(f_j(t_k+\tau))\right)\\
&-\frac{JT}{4}-2Q_{I_j(t_k+\tau)}(t_k+\tau)b_{I_j(t_k+\tau),j}(f_j(t_k+\tau))\\
\geqslant&\max_i Q_i(f_j(t_k+\tau))\mu_{i,j}(f_j(t_k+\tau))-\frac{JT}{4}-2Q_{I_j(t_k+\tau)}(t_k+\tau)b_{I_j(t_k+\tau),j}(f_j(t_k+\tau)),
\end{aligned} \tag{36}$$

where the last inequality is due to (33) and the definition of the concentration event $\mathcal{E}_{t_k,j}$ in (10). By (33) and Lemma 5, the term $Q_i(f_j(t_k+\tau))\mu_{i,j}(f_j(t_k+\tau))$ in (36) can be bounded by

$$Q_i(f_j(t_k+\tau))\mu_{i,j}(f_j(t_k+\tau)) \geqslant [Q_i(t_k)-2JT]\,\mu_{i,j}(f_j(t_k+\tau)) \geqslant Q_i(t_k)\mu_{i,j}(f_j(t_k+\tau))-2JT, \tag{37}$$

where the last inequality holds since $\mu_{i,j}(f_j(t_k+\tau)) \leqslant 1$. Substituting (37) into (36) and then into (32), we have

$$\begin{aligned}
&\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\sum_i Q_i(t_k+\tau)\mathbb{1}_{i,j}(t_k+\tau)\,|\mathcal{E}_{t_k,j}\right]\\
\geqslant&\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\left(\max_i q_i\mu_{i,j}(f_j(t_k+\tau)) - \frac{9JT}{4}\right.\right.\\
&\qquad\left.\left.-2Q_{I_j(t_k+\tau)}(t_k+\tau)b_{I_j(t_k+\tau),j}(f_j(t_k+\tau))\right)\frac{\mathbb{1}_{I_j(t_k+\tau),j}(t_k+\tau)}{\mu_{I_j(t_k+\tau),j}(f_j(t_k+\tau))}\,|\mathcal{E}_{t_k,j}\right]\\
\geqslant&\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\max_i q_i\mu_{i,j}(f_j(t_k+\tau))\frac{\mathbb{1}_{I_j(t_k+\tau),j}(t_k+\tau)}{\mu_{I_j(t_k+\tau),j}(f_j(t_k+\tau))}\,|\mathcal{E}_{t_k,j}\right] - \frac{9JT^2U_{\mathrm{S}}}{4} \qquad (38)\\
&-2U_{\mathrm{S}}\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1}Q_{I_j(t_k+\tau)}(t_k+\tau)b_{I_j(t_k+\tau),j}(f_j(t_k+\tau))\mathbb{1}_{I_j(t_k+\tau),j}(t_k+\tau)\,|\mathcal{E}_{t_k,j}\right], \qquad (39)
\end{aligned}$$

where the last inequality is by Lemma 3 and the fact that $\frac{1}{\mu_{I_j(t_k+\tau),j}(f_j(t_k+\tau))} \leqslant U_{\mathrm{S}}$.

B.4.3 Step 3: Bounding the Sum of Queue-Length-Weighted UCB Bonuses

We first look at the term in (39):

$$\begin{aligned}
&\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1}Q_{I_j(t_k+\tau)}(t_k+\tau)b_{I_j(t_k+\tau),j}(f_j(t_k+\tau))\mathbb{1}_{I_j(t_k+\tau),j}(t_k+\tau)\,|\mathcal{E}_{t_k,j}\right]\\
\leqslant&\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\left(Q_{I_j(t_k+\tau)}(f_j(t_k+\tau)) + \frac{U_{\mathrm{A}}T}{8}\right)b_{I_j(t_k+\tau),j}(f_j(t_k+\tau))\mathbb{1}_{I_j(t_k+\tau),j}(t_k+\tau)\,|\mathcal{E}_{t_k,j}\right]\\
\leqslant&\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1}Q_{I_j(t_k+\tau)}(f_j(t_k+\tau))b_{I_j(t_k+\tau),j}(f_j(t_k+\tau))\mathbb{1}_{I_j(t_k+\tau),j}(t_k+\tau)\,|\mathcal{E}_{t_k,j}\right] + \frac{U_{\mathrm{A}}T^2}{8}, \qquad (40)
\end{aligned}$$

where the first inequality is by Lemma 5 and the fact that $t_k+\tau-f_j(t_k+\tau) \leqslant U_{\mathrm{S}} \leqslant \frac{T}{8}$, and the second inequality is due to Lemma 3 and the fact that $b_{I_j(t_k+\tau),j}(f_j(t_k+\tau)) \leqslant 1$. Note that if server j is idling in time slot $f_j(t_k+\tau)$, then $\tilde{Q}_{I_j(t_k+\tau)}(f_j(t_k+\tau)) = 0$. Hence, we have $\tilde{Q}_{I_j(t_k+\tau)}(f_j(t_k+\tau)) = \tilde{Q}_{I_j(t_k+\tau)}(f_j(t_k+\tau))\eta_j(f_j(t_k+\tau))$. Also note that $0 \leqslant Q_{I_j(t_k+\tau)}(f_j(t_k+\tau)) - \tilde{Q}_{I_j(t_k+\tau)}(f_j(t_k+\tau)) \leqslant J$ by definition. Hence, we have

$$\begin{aligned}
Q_{I_j(t_k+\tau)}(f_j(t_k+\tau)) \leqslant&\tilde{Q}_{I_j(t_k+\tau)}(f_j(t_k+\tau))\eta_j(f_j(t_k+\tau)) + J\\
\leqslant&Q_{I_j(t_k+\tau)}(f_j(t_k+\tau))\eta_j(f_j(t_k+\tau)) + J. \qquad (41)
\end{aligned}$$

Substituting the bound (41) into (40) and using Lemma 3, we have

$$\begin{aligned}
&\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1}Q_{I_j(t_k+\tau)}(t_k+\tau)b_{I_j(t_k+\tau),j}(f_j(t_k+\tau))\mathbb{1}_{I_j(t_k+\tau),j}(t_k+\tau)\,|\mathcal{E}_{t_k,j}\right]\\
\leqslant&\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1}Q_{I_j(t_k+\tau)}(f_j(t_k+\tau))b_{I_j(t_k+\tau),j}(f_j(t_k+\tau))\eta_j(f_j(t_k+\tau))\mathbb{1}_{I_j(t_k+\tau),j}(t_k+\tau)\,|\mathcal{E}_{t_k,j}\right]\\
&+JT + \frac{U_{\mathrm{A}}T^2}{8}. \qquad (42)
\end{aligned}$$

By (33) and Lemma 5, we have $Q_{I_j(t_k+\tau)}(f_j(t_k+\tau)) \leqslant Q_{I_j(t_k+\tau)}(t_k) + 2TU_{\mathrm{A}} \leqslant \sum_i Q_i(t_k) + 2TU_{\mathrm{A}}$. Substituting this bound into (42) and using Lemma 3, we have

$$\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1}Q_{I_j(t_k+\tau)}(t_k+\tau)b_{I_j(t_k+\tau),j}(f_j(t_k+\tau))\mathbb{1}_{I_j(t_k+\tau),j}(t_k+\tau)\,|\mathcal{E}_{t_k,j}\right]$$

$$\leqslant\left(\sum_i q_i\right)\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1} b_{I_j(t_k+\tau),j}(f_j(t_k+\tau))\eta_j(f_j(t_k+\tau))\mathbb{1}_{I_j(t_k+\tau),j}(t_k+\tau)\,|\mathcal{E}_{t_k,j}\right]$$
$$+2T^2U_{\mathrm{A}}+JT+\frac{U_{\mathrm{A}}T^2}{8}. \tag{43}$$

Hence, by Lemma 2, (43), and (39), we have

$$(39)\geqslant-(198IU_{\mathrm{S}}^2\sqrt{T}\log T)\sum_i q_i-4T^2U_{\mathrm{S}}U_{\mathrm{A}}-2U_{\mathrm{S}}JT-\frac{U_{\mathrm{S}}U_{\mathrm{A}}T^2}{4}. \tag{44}$$

B.4.4 Step 4: Bounding the Weighted Sum of Job Completion Indicators

We next look at the term in (38):

$$\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\max_i q_i\mu_{i,j}(f_j(t_k+\tau))\frac{\mathbb{1}_{I_j(t_k+\tau),j}(t_k+\tau)}{\mu_{I_j(t_k+\tau),j}(f_j(t_k+\tau))}\,|\mathcal{E}_{t_k,j}\right]$$

Let $v_j(t)\coloneqq\max_i q_i\mu_{i,j}(t)$ for any time slot t. By law of total expectation, we have

$$\begin{aligned}
&\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1} v_j(f_j(t_k+\tau))\frac{\mathbb{1}_{I_j(t_k+\tau),j}(t_k+\tau)}{\mu_{I_j(t_k+\tau),j}(f_j(t_k+\tau))}\,|\mathcal{E}_{t_k,j}\right]\\
\geqslant&\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\frac{v_j(f_j(t_k+\tau))\mathbb{1}_{I_j(t_k+\tau),j}(t_k+\tau)}{\mu_{I_j(t_k+\tau),j}(f_j(t_k+\tau))}\right]\\
&-\hat{P}_{t_k}(\mathcal{E}_{t_k,j}^{\mathrm{c}})\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\frac{v_j(f_j(t_k+\tau))\mathbb{1}_{I_j(t_k+\tau),j}(t_k+\tau)}{\mu_{I_j(t_k+\tau),j}(f_j(t_k+\tau))}\,|\mathcal{E}_{t_k,j}^{\mathrm{c}}\right]\\
\geqslant&\sum_{i=1}^{I}\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\frac{v_j(f_j(t_k+\tau))\mathbb{1}_{i,j}(t_k+\tau)\mathbb{1}_{I_j(t_k+\tau)=i}}{\mu_{i,j}(f_j(t_k+\tau))}\right]-\frac{D_{k+1}IU_{\mathrm{S}}\sum_i q_i}{T^2},
\end{aligned}\tag{45}$$

where the last inequality is by Lemma 1 and the facts that $v_j(t)\leqslant\max_i q_i\leqslant\sum_i q_i$ and $\frac{1}{\mu_{i,j}(t)}\leqslant U_{\mathrm{S}}$ for any i,j,t. We can write the first term of (45) in a different form by summing over the time slots in which the jobs start, i.e.,

$$\begin{aligned}
&\sum_{i=1}^{I}\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\frac{v_j(f_j(t_k+\tau))\mathbb{1}_{i,j}(t_k+\tau)\mathbb{1}_{I_j(t_k+\tau)=i}}{\mu_{i,j}(f_j(t_k+\tau))}\right]\\
\geqslant&\sum_{i=1}^{I}\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\frac{v_j(t_k+\tau)\mathbb{1}_{\hat{i}_j^*(t_k+\tau)=i}}{\mu_{i,j}(t_k+\tau)}\right]-U_{\mathrm{S}}\sum_i q_i,
\end{aligned}\tag{46}$$

where the inequality holds since the last job starting before $t_k+D_k+D_{k+1}$ may not finish before $t_k+D_k+D_{k+1}$ and the first job finishing at or after t_k+D_k may not start at or after t_k+D_k, and we also use the fact that $v_j(t)\leqslant\sum_i q_i$ and $1/\mu_{i,j}(t)\leqslant U_{\mathrm{S}}$ for all t. Define $X_{i,j}(t)$ such that

$$X_{i,j}(t)\coloneqq\begin{cases}S_{i,j}(t), & \text{if } \eta_j(t)=1 \text{ (not idling);}\\ 1, & \text{if } \eta_j(t)=0 \text{ (idling).}\end{cases}$$

Note that $1/\mu_{i,j}(t_k+\tau)=E[S_{i,j}(t_k+\tau)]\geqslant E[X_{i,j}(t_k+\tau)]$. Then we have

$$\begin{aligned}
&\sum_{i=1}^{I}\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\frac{v_j(t_k+\tau)\mathbb{1}_{\hat{i}_j^*(t_k+\tau)=i}}{\mu_{i,j}(t_k+\tau)}\right]\\
\geqslant&\sum_{i=1}^{I}\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1}v_j(t_k+\tau)\mathbb{1}_{\hat{i}_j^*(t_k+\tau)=i}E[X_{i,j}(t_k+\tau)]\right].
\end{aligned}\tag{47}$$

From (46) and (47), we have

$$\begin{aligned}
&\sum_{i=1}^{I} \hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1} \frac{v_j(f_j(t_k+\tau))\mathbb{1}_{i,j}(t_k+\tau)\mathbb{1}_{I_j(t_k+\tau)=i}}{\mu_{i,j}(f_j(t_k+\tau))}\right] \\
&\geqslant \sum_{i=1}^{I} \hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1} v_j(t_k+\tau)\mathbb{1}_{\hat{i}_j^*(t_k+\tau)=i} E[X_{i,j}(t_k+\tau)]\right] - U_{\mathrm{S}}\sum_i q_i \\
&= \sum_{i=1}^{I} \hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1} v_j(t_k+\tau)\mathbb{1}_{\hat{i}_j^*(t_k+\tau)=i}\left(E[X_{i,j}(t_k+\tau)] - X_{i,j}(t_k+\tau)\right)\right] \\
&\quad + \sum_{i=1}^{I} \hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1} v_j(t_k+\tau)\mathbb{1}_{\hat{i}_j^*(t_k+\tau)=i} X_{i,j}(t_k+\tau)\right] - U_{\mathrm{S}}\sum_i q_i. \qquad (48)
\end{aligned}$$

Note that the term $\sum_{i=1}^{I} \hat{E}_{t_k}[\sum_{\tau=D_k}^{D_k+D_{k+1}-1} v_j(t_k+\tau)\mathbb{1}_{\hat{i}_j^*(t_k+\tau)=i} X_{i,j}(t_k+\tau)]$ in (48) can be rewritten using f_j in the following way:

$$\sum_{i=1}^{I} \hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1} v_j(t_k+\tau)\mathbb{1}_{\hat{i}_j^*(t_k+\tau)=i} X_{i,j}(t_k+\tau)\right] = \hat{E}_{t_k}\left[\sum_{\tau=\tau_{\mathrm{start}}}^{\tau_{\mathrm{end}}} v_j(f_j(t_k+\tau))\right], \qquad (49)$$

where $t_k+\tau_{\mathrm{start}}$ is the starting (or idling) time of the first schedule that starts at or after t_k+D_k and $t_k+\tau_{\mathrm{end}}$ is the finishing (or idling) time of the last schedule that starts at or before $t_k+D_k+D_{k+1}-1$. By the facts that $t_k+\tau_{\mathrm{start}} < t_k+D_k+U_{\mathrm{S}}$, $t_k+\tau_{\mathrm{end}} \geqslant t_k+D_k+D_{k+1}-1$, and $v_j(t) \leqslant \sum_i q_i$, we have

$$\begin{aligned}
&\hat{E}_{t_k}\left[\sum_{\tau=t_{\mathrm{start}}}^{t_{\mathrm{end}}} v_j(f_j(t_k+\tau))\right] \geqslant \hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1} v_j(f_j(t_k+\tau))\right] - U_{\mathrm{S}}\sum_i q_i \\
&\geqslant \sum_{\tau=D_k}^{D_k+D_{k+1}-1} \max_i q_i\left(\mu_{i,j}(t_k+\tau) - \frac{1}{T^p}\right) - U_{\mathrm{S}}\sum_i q_i \\
&\geqslant \sum_{\tau=D_k}^{D_k+D_{k+1}-1} \max_i q_i\mu_{i,j}(t_k+\tau) - \left(\frac{D_{k+1}}{T^p} + U_{\mathrm{S}}\right)\sum_i q_i, \qquad (50)
\end{aligned}$$

where the first inequality uses the fact that $t_k+\tau-f_j(t_k+\tau) \leqslant U_{\mathrm{S}}$ and the second condition in Assumption 1. Then, combining (45), (48), (49), and (50), we have

$$\begin{aligned}
&\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1} \frac{v_j(f_j(t_k+\tau))\mathbb{1}_{I_j(t_k+\tau),j}(t_k+\tau)}{\mu_{I_j(t_k+\tau),j}(f_j(t_k+\tau))}\,|\mathcal{E}_{t_k,j}\right] \\
&\geqslant \sum_{i=1}^{I} \hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1} v_j(t_k+\tau)\mathbb{1}_{\hat{i}_j^*(t_k+\tau)=i}\left(E[X_{i,j}(t_k+\tau)] - X_{i,j}(t_k+\tau)\right)\right] \\
&\quad + \sum_{\tau=D_k}^{D_k+D_{k+1}-1} \max_i q_i\mu_{i,j}(t_k+\tau) - \left(\frac{D_{k+1}}{T^p} + 2U_{\mathrm{S}} + \frac{D_{k+1}IU_{\mathrm{S}}}{T^2}\right)\sum_i q_i. \qquad (51)
\end{aligned}$$

Next let us look at the first term of (51). Note that the differences $E[X_{i,j}(t_k+\tau)] - X_{i,j}(t_k+\tau)$ at the idling time slots are zero by definition. Hence,

$$\begin{aligned}
&\sum_{i=1}^{I} \hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1} v_j(t_k+\tau)\mathbb{1}_{\hat{i}_j^*(t_k+\tau)=i}\left(E[X_{i,j}(t_k+\tau)] - X_{i,j}(t_k+\tau)\right)\right] \\
&= \sum_{i=1}^{I} \hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1} v_j(t_k+\tau)\mathbb{1}_{\hat{i}_j^*(t_k+\tau)=i}\eta_j(t_k+\tau)\left(E[X_{i,j}(t_k+\tau)] - X_{i,j}(t_k+\tau)\right)\right]. \qquad (52)
\end{aligned}$$

Consider the following concentration event

$$\mathcal{E}_{X,t_k,i,j} := \left\{ \sum_{\tau=D_k}^{D_k+D_{k+1}-1} v_j(t_k+\tau)\mathbb{1}_{i_j^*(t_k+\tau)=i}\eta_j(t_k+\tau)\left(E[X_{i,j}(t_k+\tau)] - X_{i,j}(t_k+\tau)\right) \geqslant -U_{\mathrm{S}}\sqrt{2T\log T}\sum_i q_i \right\}. \quad (53)$$

We have the following lemma:

Lemma 7. *For any $k \geqslant 0$, $i \in \{1,\dots,I\}$, $j \in \{1,\dots,J\}$, we have $\hat{P}_{t_k}\left(\mathcal{E}^{\mathrm{c}}_{X,t_k,i,j}\right) \leqslant \frac{1}{T^2}$.*

Proof of this lemma can be found in Section C.7. Then by law of total expectation, definition (53), and the bounds on service times and $v_j(t)$, we have

$$\begin{aligned}
&\sum_{i=1}^{I}\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1} v_j(t_k+\tau)\mathbb{1}_{i_j^*(t_k+\tau)=i}\eta_j(t_k+\tau)\left(E[X_{i,j}(t_k+\tau)] - X_{i,j}(t_k+\tau)\right)\right]\\
\geqslant&\sum_{i=1}^{I}\left[\hat{P}_{t_k}\left(\mathcal{E}_{X,t_k,i,j}\right)\left(-U_{\mathrm{S}}\sqrt{2T\log T}\sum_i q_i\right) + \hat{P}_{t_k}\left(\mathcal{E}^{\mathrm{c}}_{X,t_k,i,j}\right)\left(-D_{k+1}U_{\mathrm{S}}\sum_i q_i\right)\right]\\
\geqslant&-\left(IU_{\mathrm{S}}\sqrt{2T\log T} + \frac{D_{k+1}IU_{\mathrm{S}}}{T^2}\right)\sum_i q_i, \qquad (54)
\end{aligned}$$

where the last inequality is by Lemma 7. Combining (38), (51), (52), and (54), we have

$$\begin{aligned}
(38) =&\ \hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1} v_j(f_j(t_k+\tau))\frac{\mathbb{1}_{I_j(t_k+\tau),j}(t_k+\tau)}{\mu_{I_j(t_k+\tau),j}(f_j(t_k+\tau))}\,|\mathcal{E}_{t_k,j}\right] - \frac{9JT^2U_{\mathrm{S}}}{4}\\
\geqslant&\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\max_i q_i\mu_{i,j}(t_k+\tau) - \left(\frac{D_{k+1}}{T^p} + 2U_{\mathrm{S}} + IU_{\mathrm{S}}\sqrt{2T\log T} + \frac{2D_{k+1}IU_{\mathrm{S}}}{T^2}\right)\sum_i q_i - \frac{9JT^2U_{\mathrm{S}}}{4}. \qquad (55)
\end{aligned}$$

Combining (38), (39), (55), and (44) and using Lemma 6 on $\sum_i q_i$ and Lemma 3 on D_{k+1}, we have

$$\begin{aligned}
&\hat{E}_{t_k}\left[\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\sum_i Q_i(t_k+\tau)\mathbb{1}_{i,j}(t_k+\tau)\,|\mathcal{E}_{t_k,j}\right]\\
\geqslant&\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\max_i q_i\mu_{i,j}(t_k+\tau)\\
&-\left(\frac{D_{k+1}}{T^p} + 2U_{\mathrm{S}} + IU_{\mathrm{S}}\sqrt{2T\log T} + \frac{2D_{k+1}IU_{\mathrm{S}}}{T^2} + 198IU_{\mathrm{S}}^2\sqrt{T}\log T\right)\sum_i q_i\\
&-\frac{9JT^2U_{\mathrm{S}}}{4} - 4T^2U_{\mathrm{S}}U_{\mathrm{A}} - 2U_{\mathrm{S}}JT - \frac{U_{\mathrm{S}}U_{\mathrm{A}}T^2}{4}\\
\geqslant&\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\max_i q_i\mu_{i,j}(t_k+\tau) - \frac{401IU_{\mathrm{S}}^2\log T}{T^{\min\{\frac{1}{2},p\}}}\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\hat{E}_{t_k}\left[\sum_i Q_i(t_k+\tau)\right] - 407IJU_{\mathrm{S}}^2U_{\mathrm{A}}T^2. \qquad (56)
\end{aligned}$$

Substituting (56) into (31), we have

$$\begin{aligned}
&\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\hat{E}_{t_k}\left[\sum_i Q_i(t_k+\tau)\mathbb{1}_{i,j}(t_k+\tau)\right]\\
\geqslant&\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\max_i q_i\mu_{i,j}(t_k+\tau) - \frac{402IU_{\mathrm{S}}^2\log T}{T^{\min\{\frac{1}{2},p\}}}\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\hat{E}_{t_k}\left[\sum_i Q_i(t_k+\tau)\right] - 408I^2JU_{\mathrm{S}}^2U_{\mathrm{A}}T^2. \qquad (57)
\end{aligned}$$

Substituting (57) into (20), we finally obtain the bound for the service term (20):

$$(20) \leqslant -2\sum_j\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\max_i q_i\mu_{i,j}(t_k+\tau)$$

$$+\frac{804IJU_{\mathrm{S}}^2\log T}{T^{\min\{\frac{1}{2},p\}}}\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\hat{E}_{t_k}\left[\sum_i Q_i(t_k+\tau)\right]+817I^2J^2U_{\mathrm{S}}^2U_{\mathrm{A}}^2T^2. \tag{58}$$

In the next subsection, we will combine the bounds of the arrival term and the service term and then sum over all intervals.

B.5 Telescoping Sum

Combining (19), (20), (27), and (58), we have

$$\begin{aligned}
&E\left[L(t_k+D_k+D_{k+1})-L(t_k+D_k)\,|\boldsymbol{Q}(t_k)=\boldsymbol{q},\boldsymbol{H}(t_k)=\boldsymbol{h}\right]\\
\leqslant&\left(\frac{804IJU_{\mathrm{S}}^2\log T}{T^{\min\{\frac{1}{2},p\}}}-2\delta\right)\sum_{\tau=D_k}^{D_k+D_{k+1}-1}\hat{E}_{t_k}\left[\sum_i Q_i(t_k+\tau)\right]+821I^2J^2U_{\mathrm{S}}^2U_{\mathrm{A}}^2T^2\\
\leqslant&-\delta\sum_{\tau=D_k}^{D_k+D_{k+1}-1}E\left[\sum_i Q_i(t_k+\tau)\,|\boldsymbol{Q}(t_k)=\boldsymbol{q},\boldsymbol{H}(t_k)=\boldsymbol{h}\right]+821I^2J^2U_{\mathrm{S}}^2U_{\mathrm{A}}^2T^2
\end{aligned}$$

since $\frac{804IJU_{\mathrm{S}}^2\log T}{T^{\min\{\frac{1}{2},p\}}}\leqslant\delta$. Taking expectation on both sides, we have

$$E\left[L(t_k+D_k+D_{k+1})-L(t_k+D_k)\right]\leqslant-\delta\sum_{\tau=D_k}^{D_k+D_{k+1}-1}E\left[\sum_i Q_i(t_k+\tau)\right]+821I^2J^2U_{\mathrm{S}}^2U_{\mathrm{A}}^2T^2. \tag{59}$$

Let $t\geqslant T$. Since $D_0\leqslant T\leqslant t$ (by Lemma 3), we have

$$\frac{1}{t}\sum_{\tau=1}^{t}E\left[\sum_i Q_i(\tau)\right]=\frac{1}{t}\sum_{\tau=1}^{D_0-1}E\left[\sum_i Q_i(\tau)\right]+\frac{1}{t}\sum_{\tau=D_0}^{t}E\left[\sum_i Q_i(\tau)\right].$$

Note that there exists an integer K such that $t\leqslant\sum_{k=0}^{K}D_k-1<t+\frac{T}{2}+W$ by Lemma 3. Then we have

$$\sum_{k=0}^{K}D_k-1\geqslant t\geqslant\sum_{k=0}^{K}D_k-\frac{T}{2}-W\geqslant\sum_{k=1}^{K}D_k-W, \tag{60}$$

where the last inequality is by Lemma 3. Hence, we have

$$\begin{aligned}
\frac{1}{t}\sum_{\tau=1}^{t}E\left[\sum_i Q_i(\tau)\right]=&\frac{1}{t}\sum_{\tau=1}^{D_0-1}E\left[\sum_i Q_i(\tau)\right]+\frac{\sum_{k=1}^{K}D_k}{t}\frac{1}{\sum_{k=1}^{K}D_k}\sum_{\tau=D_0}^{t}E\left[\sum_i Q_i(\tau)\right]\\
\leqslant&\frac{1}{t}\sum_{\tau=1}^{D_0-1}E\left[\sum_i Q_i(\tau)\right]+\frac{t+W}{t}\frac{1}{\sum_{k=1}^{K}D_k}\sum_{\tau=D_0}^{\sum_{k=0}^{K}D_k-1}E\left[\sum_i Q_i(\tau)\right].
\end{aligned} \tag{61}$$

Summing both sides of (59) over $k=0,1,\ldots,K-1$, we have

$$E\left[L\left(\sum_{k=0}^{K}D_k\right)-L\left(D_0\right)\right]\leqslant-\delta\sum_{\tau=D_0}^{\sum_{k=0}^{K}D_k-1}E\left[\sum_i Q_i(\tau)\right]+821KI^2J^2U_{\mathrm{S}}^2U_{\mathrm{A}}^2T^2.$$

Hence, we have

$$\sum_{\tau=D_0}^{\sum_{k=0}^{K}D_k-1}E\left[\sum_i Q_i(\tau)\right]\leqslant\frac{1}{\delta}E\left[-L\left(\sum_{k=0}^{K}D_k\right)+L\left(D_0\right)\right]+\frac{821KI^2J^2U_{\mathrm{S}}^2U_{\mathrm{A}}^2T^2}{\delta}$$

Dividing both sides by $\sum_{k=1}^{K}D_k$, we have

$$\frac{1}{\sum_{k=1}^{K}D_k}\sum_{\tau=D_0}^{\sum_{k=0}^{K}D_k-1}E\left[\sum_i Q_i(\tau)\right]\leqslant\frac{1}{\delta\sum_{k=1}^{K}D_k}E\left[-L\left(\sum_{k=0}^{K}D_k\right)+L\left(D_0\right)\right]+\frac{821KI^2J^2U_{\mathrm{S}}^2U_{\mathrm{A}}^2T^2}{\delta\sum_{k=1}^{K}D_k}$$

$$\leqslant \frac{1}{\delta \sum_{k=1}^{K} D_k} E\left[L\left(D_0\right)\right] + \frac{1642 I^2 J^2 U_{\mathrm{S}}^2 U_{\mathrm{A}}^2 T}{\delta}, \tag{62}$$

where the last inequality uses Lemma 3. Substituting (62) into (61), we have

$$\begin{aligned}
\frac{1}{t}\sum_{\tau=1}^{t} E\left[\sum_i Q_i(\tau)\right] \leqslant & \frac{1}{t}\sum_{\tau=1}^{D_0-1} E\left[\sum_i Q_i(\tau)\right] + \frac{t+W}{t}\left(\frac{1}{\delta \sum_{k=1}^{K} D_k} E\left[L\left(D_0\right)\right] + \frac{1642 I^2 J^2 U_{\mathrm{S}}^2 U_{\mathrm{A}}^2 T}{\delta}\right) \\
\leqslant & \frac{I T^2 U_{\mathrm{A}}}{t} + \left(1+\frac{W}{t}\right)\left(\frac{I T^2 U_{\mathrm{A}}^2}{\delta \sum_{k=1}^{K} D_k} + \frac{1642 I^2 J^2 U_{\mathrm{S}}^2 U_{\mathrm{A}}^2 T}{\delta}\right) \\
\leqslant & \frac{I T^2 U_{\mathrm{A}}}{t} + \left(1+\frac{W}{t}\right)\left(\frac{I T^2 U_{\mathrm{A}}^2}{\delta(t+1-T)} + \frac{1642 I^2 J^2 U_{\mathrm{S}}^2 U_{\mathrm{A}}^2 T}{\delta}\right),
\end{aligned}$$

where the second inequality is obtained by using Lemma 3 and Lemma 5 to bound $Q_i(\tau)$ and $L(D_0)$ with the initial condition $Q_i(0) = 0$, and the last inequality holds since $\sum_{k=1}^{K} D_k = \sum_{k=0}^{K} D_k - D_0 \geqslant t + 1 - D_0 \geqslant t + 1 - T$ by (60) and Lemma 3. The finite-round bound in Theorem 1 is proved.

Letting $t \to \infty$, we obtain

$$\limsup_{t\to\infty} \frac{1}{t}\sum_{\tau=1}^{t} E\left[\sum_i Q_i(\tau)\right] \leqslant \frac{1642 I^2 J^2 U_{\mathrm{S}}^2 U_{\mathrm{A}}^2 T}{\delta}.$$

The asymptotic bound in Theorem 1 is proved.

C PROOFS OF ALL LEMMAS

In this section, we present the proofs of all the lemmas that appeared in the paper.

C.1 Proof of Lemma 1

Lemma 1. *For any $k \geqslant 0$, we have*

$$\hat{P}_{t_k}(\mathcal{E}_{t_k,j}^{\mathrm{c}}) \coloneqq \Pr(\mathcal{E}_{t_k,j}^{\mathrm{c}} \,|\mathbf{Q}(t_k) = \mathbf{q}, \mathbf{H}(t_k) = \mathbf{h}) \leqslant \frac{I}{T^2},$$

where

$$\mathcal{E}_{t_k,j} \coloneqq \left\{ \textit{for all } i, \tau \in \left[D_k - \frac{T}{8}, D_k + D_{k+1} - 1\right], |\hat{\mu}_{i,j}(t_k+\tau) - \mu_{i,j}(t_k+\tau)| \leqslant b_{i,j}(t_k+\tau) \right\}.$$

Proof. We define another event as follows:

$$\mathcal{E}'_{t_k,j} \coloneqq \left\{ \text{for all } i, \tau \in \left[D_k - \frac{T}{8}, D_k + D_{k+1} - 1\right], \left|\frac{1}{\hat{\mu}_{i,j}(t_k+\tau)} - \frac{1}{\mu_{i,j}(t_k+\tau)}\right| \leqslant \sqrt{\frac{4U_{\mathrm{S}}^2 \log T}{\hat{N}_{i,j}(t_k+\tau)}} \right\}.$$

We first show that $\mathcal{E}'_{t_k,j} \subseteq \mathcal{E}_{t_k,j}$. Suppose we have

$$\left|\frac{1}{\hat{\mu}_{i,j}(t_k+\tau)} - \frac{1}{\mu_{i,j}(t_k+\tau)}\right| \leqslant \sqrt{\frac{4U_{\mathrm{S}}^2 \log T}{\hat{N}_{i,j}(t_k+\tau)}}.$$

Then

$$\frac{1}{\hat{\mu}_{i,j}(t_k+\tau)} - \sqrt{\frac{4U_{\mathrm{S}}^2 \log T}{\hat{N}_{i,j}(t_k+\tau)}} \leqslant \frac{1}{\mu_{i,j}(t_k+\tau)} \leqslant \frac{1}{\hat{\mu}_{i,j}(t_k+\tau)} + \sqrt{\frac{4U_{\mathrm{S}}^2 \log T}{\hat{N}_{i,j}(t_k+\tau)}},$$

which implies that

$$\frac{1}{\frac{1}{\hat{\mu}_{i,j}(t_k+\tau)}+\sqrt{\frac{4U_{\mathrm{S}}^2\log T}{\hat{N}_{i,j}(t_k+\tau)}}}\leqslant\mu_{i,j}(t_k+\tau)\leqslant\frac{1}{\max\left\{\frac{1}{\hat{\mu}_{i,j}(t_k+\tau)}-\sqrt{\frac{4U_{\mathrm{S}}^2\log T}{\hat{N}_{i,j}(t_k+\tau)}},1\right\}}$$

due to the fact that $\mu_{i,j}(t_k+\tau)\leqslant 1$. Note that we also have

$$\frac{1}{\frac{1}{\hat{\mu}_{i,j}(t_k+\tau)}+\sqrt{\frac{4U_{\mathrm{S}}^2\log T}{\hat{N}_{i,j}(t_k+\tau)}}}\leqslant\frac{1}{\frac{1}{\hat{\mu}_{i,j}(t_k+\tau)}}=\hat{\mu}_{i,j}(t_k+\tau)\leqslant\frac{1}{\max\left\{\frac{1}{\hat{\mu}_{i,j}(t_k+\tau)}-\sqrt{\frac{4U_{\mathrm{S}}^2\log T}{\hat{N}_{i,j}(t_k+\tau)}},1\right\}}$$

due to the fact that $\hat{\mu}_{i,j}(t_k+\tau)\leqslant 1$. Hence,

$$\begin{aligned}
|\hat{\mu}_{i,j}(t_k+\tau)-\mu_{i,j}(t_k+\tau)|\leqslant&\frac{1}{\max\left\{\frac{1}{\hat{\mu}_{i,j}(t_k+\tau)}-\sqrt{\frac{4U_{\mathrm{S}}^2\log T}{\hat{N}_{i,j}(t_k+\tau)}},1\right\}}-\frac{1}{\frac{1}{\hat{\mu}_{i,j}(t_k+\tau)}+\sqrt{\frac{4U_{\mathrm{S}}^2\log T}{\hat{N}_{i,j}(t_k+\tau)}}}\\
=&\frac{\frac{1}{\hat{\mu}_{i,j}(t_k+\tau)}+\sqrt{\frac{4U_{\mathrm{S}}^2\log T}{\hat{N}_{i,j}(t_k+\tau)}}-\max\left\{\frac{1}{\hat{\mu}_{i,j}(t_k+\tau)}-\sqrt{\frac{4U_{\mathrm{S}}^2\log T}{\hat{N}_{i,j}(t_k+\tau)}},1\right\}}{\max\left\{\frac{1}{\hat{\mu}_{i,j}(t_k+\tau)}-\sqrt{\frac{4U_{\mathrm{S}}^2\log T}{\hat{N}_{i,j}(t_k+\tau)}},1\right\}\left(\frac{1}{\hat{\mu}_{i,j}(t_k+\tau)}+\sqrt{\frac{4U_{\mathrm{S}}^2\log T}{\hat{N}_{i,j}(t_k+\tau)}}\right)}\\
\leqslant&\frac{2\sqrt{\frac{4U_{\mathrm{S}}^2\log T}{\hat{N}_{i,j}(t_k+\tau)}}}{\max\left\{\frac{1}{\hat{\mu}_{i,j}(t_k+\tau)}-\sqrt{\frac{4U_{\mathrm{S}}^2\log T}{\hat{N}_{i,j}(t_k+\tau)}},1\right\}\left(\frac{1}{\hat{\mu}_{i,j}(t_k+\tau)}+\sqrt{\frac{4U_{\mathrm{S}}^2\log T}{\hat{N}_{i,j}(t_k+\tau)}}\right)}\\
\leqslant&2\sqrt{\frac{4U_{\mathrm{S}}^2\log T}{\hat{N}_{i,j}(t_k+\tau)}},
\end{aligned}$$

where the last inequality is due to the fact that $\frac{1}{\hat{\mu}_{i,j}(t_k+\tau)}\geqslant 1$. Hence, we have

$$|\hat{\mu}_{i,j}(t_k+\tau)-\mu_{i,j}(t_k+\tau)|\leqslant\min\left\{2\sqrt{\frac{4U_{\mathrm{S}}^2\log T}{\hat{N}_{i,j}(t_k+\tau)}},1\right\}=b_{i,j}(t_k+\tau)$$

since $\hat{\mu}_{i,j}(t_k+\tau),\mu_{i,j}(t_k+\tau)\in[0,1]$. $\mathcal{E}'_{t_k,j}\subseteq\mathcal{E}_{t_k,j}$ is proved.

Next, it remains to show that

$$\hat{P}_{t_k}(\mathcal{E}'^{\mathrm{c}}_{t_k,j})\leqslant\frac{I}{T^2}.$$

Consider the event

$$\mathcal{E}'_{t_k,i,j,\tau}:=\left\{\left|\frac{1}{\hat{\mu}_{i,j}(t_k+\tau)}-\frac{1}{\mu_{i,j}(t_k+\tau)}\right|\leqslant\sqrt{\frac{4U_{\mathrm{S}}^2\log T}{\hat{N}_{i,j}(t_k+\tau)}}\right\}.$$

We have

$$\begin{aligned}
\hat{P}_{t_k}\left(\mathcal{E}'^{\mathrm{c}}_{t_k,i,j,\tau}\right)=&\hat{P}_{t_k}\left(\left|\frac{1}{\hat{\mu}_{i,j}(t_k+\tau)}-\frac{1}{\mu_{i,j}(t_k+\tau)}\right|>\sqrt{\frac{4U_{\mathrm{S}}^2\log T}{\hat{N}_{i,j}(t_k+\tau)}}\right)\\
=&\hat{P}_{t_k}\left(\left|\frac{1}{\hat{\mu}_{i,j}(t_k+\tau)}-\frac{1}{\mu_{i,j}(t_k+\tau)}\right|>\sqrt{\frac{4U_{\mathrm{S}}^2\log T}{\hat{N}_{i,j}(t_k+\tau)}},\hat{N}_{i,j}(t_k+\tau)>4\log T\right)\\
&+\hat{P}_{t_k}\left(\left|\frac{1}{\hat{\mu}_{i,j}(t_k+\tau)}-\frac{1}{\mu_{i,j}(t_k+\tau)}\right|>\sqrt{\frac{4U_{\mathrm{S}}^2\log T}{\hat{N}_{i,j}(t_k+\tau)}},\hat{N}_{i,j}(t_k+\tau)\leqslant 4\log T\right).
\end{aligned}$$

Since

$$\begin{aligned}&\hat{P}_{t_k}\left(\left|\frac{1}{\hat{\mu}_{i,j}(t_k+\tau)}-\frac{1}{\mu_{i,j}(t_k+\tau)}\right|>\sqrt{\frac{4U_{\mathrm{S}}^2\log T}{\hat{N}_{i,j}(t_k+\tau)}},\hat{N}_{i,j}(t_k+\tau)\leqslant 4\log T\right)\\ \leqslant&\hat{P}_{t_k}\left(\left|\frac{1}{\hat{\mu}_{i,j}(t_k+\tau)}-\frac{1}{\mu_{i,j}(t_k+\tau)}\right|>U_{\mathrm{S}}\right)=0,\end{aligned}$$

we then have

$$\begin{aligned}&\hat{P}_{t_k}\left(\mathcal{E}'^{\mathrm{c}}_{t_k,i,j,\tau}\right)\\ =&\hat{P}_{t_k}\left(\left|\frac{1}{\hat{\mu}_{i,j}(t_k+\tau)}-\frac{1}{\mu_{i,j}(t_k+\tau)}\right|>\sqrt{\frac{4U_{\mathrm{S}}^2\log T}{\hat{N}_{i,j}(t_k+\tau)}},\hat{N}_{i,j}(t_k+\tau)>4\log T\right)\\ =&\hat{P}_{t_k}\left(\left|\hat{\phi}_{i,j}(t_k+\tau)-\frac{\hat{N}_{i,j}(t_k+\tau)}{\mu_{i,j}(t_k+\tau)}\right|>\sqrt{4\hat{N}_{i,j}(t_k+\tau)U_{\mathrm{S}}^2\log T},\hat{N}_{i,j}(t_k+\tau)>4\log T\right).\end{aligned}\tag{63}$$

Define a random mapping f_j that maps a time slot to another time slot such that if $y=f_j(x)$ then y is the time slot when server j picked the job that was being served at server j in time slot x. If server j was idling in time slot x, then let $f_j(x)=x$. That is,

$$f_j(x)\coloneqq\max\left\{t:t\leqslant x,\hat{i}_j^*(t)=I_j(x)\right\}.$$

Let M_1 be such that t_k+M_1 is the first time slot when server j picked queue i at or after t_k, i,e.,

$$M_1\coloneqq\min\left\{m:m\geqslant 0,\hat{i}_j^*(t_k+m)=i\right\}.$$

Let $M_2\coloneqq M_1+(S_{i,j}(t_k+M_1)-1)\eta_j(t_k+M_1)$. Then t_k+M_2 is the time slot when the job picked by server j at time t_k+M_1 was completed or if server j was idling because the selected waiting queue is empty at t_k+M_1, $M_2=M_1$. Hence, $t_k+M_1=f_j(t_k+M_2)$. Note that M_1,M_2 are random variables. Then according to the algorithm, we have

$$\begin{aligned}&\hat{\phi}_{i,j}(t_k+\tau)\\ =&\gamma^{\tau-M_2}\hat{\phi}_{i,j}(t_k+M_2)\\ &+\sum_{m=M_2+1}^{\tau}\gamma^{\tau-m+M_{i,j}(t_k+m-1)}\mathbb{1}_{i,j}(t_k+m-1)\eta_j(t_k+m-1)[M_{i,j}(t_k+m-1)+1]\\ =&\gamma^{\tau-M_2}\hat{\phi}_{i,j}(t_k+M_2)\\ &+\sum_{m=M_2+1}^{\tau}\gamma^{\tau-m+S_{i,j}(f_j(t_k+m-1))-1}\mathbb{1}_{i,j}(t_k+m-1)\eta_j(f_j(t_k+m-1))S_{i,j}(f_j(t_k+m-1)),\end{aligned}$$

where the summation only includes the time slots when there is job completion of queue i at server j. This can be transformed into summing over the time slots when server j is available and picks queue i, i.e.,

$$\begin{aligned}&\hat{\phi}_{i,j}(t_k+\tau)\\ =&\gamma^{\tau-M_2}\hat{\phi}_{i,j}(t_k+M_2)+\sum_{m=M_1+1}^{M_3}\gamma^{\tau-m}\mathbb{1}_{\hat{i}_j^*(t_k+m-1)=i}\,\eta_j(t_k+m-1)S_{i,j}(t_k+m-1)\end{aligned}$$

where M_3 is a random variable such that $t_k+M_3=f_j(t_k+\tau)$. Since there is no job of type i starting at server j in the time interval $[t_k,t_k+M_1-1]$, we have

$$\begin{aligned}&\hat{\phi}_{i,j}(t_k+\tau)\\ =&\gamma^{\tau-M_2}\hat{\phi}_{i,j}(t_k+M_2)+\sum_{m=1}^{M_3}\gamma^{\tau-m}\mathbb{1}_{\hat{i}_j^*(t_k+m-1)=i}\,\eta_j(t_k+m-1)S_{i,j}(t_k+m-1).\end{aligned}$$

We claim that there is no job completion of type i at server j in the time interval $[t_k + \frac{T}{8}, t_k + M_2 - 1]$ if $M_2 - 1 \geqslant \frac{T}{8}$. We can prove it by contradiction. Suppose there is a job of type i that was completed at server j in $[t_k + \frac{T}{8}, t_k + M_2 - 1]$. Then the job must start at or after t_k since $U_{\mathrm{S}} \leqslant \frac{T}{8}$. Also, the job must start before $t_k + M_1$ since there is another job at server j starting at $t_k + M_1$ and finishes at $t_k + M_2$ by the definition of M_1 and M_2. Therefore, the job that was completed at server j in $[t_k + \frac{T}{8}, t_k + M_2 - 1]$ should start in the time interval $[t_k, t_k + M_1 - 1]$. However, by the definition of M_1, there should not be any job of type i starting at server j in $[t_k, t_k + M_1 - 1]$, which is a contradiction. Based on the claim, if $M_2 - 1 \geqslant \frac{T}{8}$, since there is no job completion, from the algorithm we know

$$\hat{\phi}_{i,j}(t_k + M_2) = \gamma^{M_2 - \frac{T}{8}} \hat{\phi}_{i,j}\left(t_k + \frac{T}{8}\right) = \gamma^{M_2 - \min\left\{\frac{T}{8}, M_2\right\}} \hat{\phi}_{i,j}\left(t_k + \min\left\{\frac{T}{8}, M_2\right\}\right)$$

where the last equality holds since $\min\left\{\frac{T}{8}, M_2\right\} = \frac{T}{8}$. If $M_2 - 1 < \frac{T}{8}$, then we have

$$\hat{\phi}_{i,j}(t_k + M_2) = \gamma^{M_2 - \min\left\{\frac{T}{8}, M_2\right\}} \hat{\phi}_{i,j}\left(t_k + \min\left\{\frac{T}{8}, M_2\right\}\right)$$

since $\min\left\{\frac{T}{8}, M_2\right\} = M_2$. Combining the above two cases, we have

$$\gamma^{\tau - M_2} \hat{\phi}_{i,j}(t_k + M_2) = \gamma^{\tau - \min\left\{\frac{T}{8}, M_2\right\}} \hat{\phi}_{i,j}\left(t_k + \min\left\{\frac{T}{8}, M_2\right\}\right).$$

We want to upper bound this term. Since

$$\hat{\phi}_{i,j}\left(t_k + \min\left\{\frac{T}{8}, M_2\right\}\right) \leqslant U_{\mathrm{S}} \sum_{t=0}^{\infty} \gamma^t = \frac{T U_{\mathrm{S}}}{8 \log T},$$

we have

$$\begin{aligned}
&\gamma^{\tau - \min\left\{\frac{T}{8}, M_2\right\}} \hat{\phi}_{i,j}\left(t_k + \min\left\{\frac{T}{8}, M_2\right\}\right) \leqslant \left(1 - \frac{8 \log T}{T}\right)^{\tau - \min\left\{\frac{T}{8}, M_2\right\}} \frac{T U_{\mathrm{S}}}{8 \log T} \\
\leqslant &\left(1 - \frac{8 \log T}{T}\right)^{D_k - \frac{T}{4}} \frac{T U_{\mathrm{S}}}{8 \log T} \leqslant \left(1 - \frac{2 \log T}{T/4}\right)^{\frac{T}{4}} \frac{T U_{\mathrm{S}}}{8 \log T} \leqslant \exp(-2 \log T) \frac{T U_{\mathrm{S}}}{8 \log T} = \frac{U_{\mathrm{S}}}{8 T \log T},
\end{aligned}$$

where the second inequality holds since $\tau \geqslant D_k - \frac{T}{8}$, the third inequality uses the bound on D_k in Lemma 3, and the last inequality is due to the fact that $\left(1 - \frac{x}{n}\right)^n \leqslant \exp(-x)$ for any $x \leqslant n$ and $n \in \mathbb{N}$. Therefore,

$$\hat{\phi}_{i,j}(t_k + \tau) \leqslant \frac{U_{\mathrm{S}}}{8 T \log T} + \sum_{m=1}^{M_3} \gamma^{\tau - m} \mathbb{1}_{\hat{i}_j^*(t_k + m - 1) = i} \, \eta_j(t_k + m - 1) S_{i,j}(t_k + m - 1). \tag{64}$$

Similarly, we have

$$\hat{N}_{i,j}(t_k + \tau) \leqslant \frac{1}{8 T \log T} + \sum_{m=1}^{M_3} \gamma^{\tau - m} \mathbb{1}_{\hat{i}_j^*(t_k + m - 1) = i} \, \eta_j(t_k + m - 1). \tag{65}$$

Note that $E[S_{i,j}(t_k + \tau)] = \frac{1}{\mu_{i,j}(t_k + \tau)}$. Let $\epsilon_m := \mathbb{1}_{\hat{i}_j^*(t_k + m - 1) = i} \eta_j(t_k + m - 1)$. Substituting (64) and (65) into (63), we have

$$\begin{aligned}
&\hat{P}_{t_k}\left(\mathcal{E}'^{\mathrm{c}}_{t_k,i,j,\tau}\right) \\
=&\hat{P}_{t_k}\left(\left|\hat{\phi}_{i,j}(t_k + \tau) - E[S_{i,j}(t_k + \tau)] \hat{N}_{i,j}(t_k + \tau)\right| > \sqrt{4 \hat{N}_{i,j}(t_k + \tau) U_{\mathrm{S}}^2 \log T}, \hat{N}_{i,j}(t_k + \tau) > 4 \log T\right) \\
\leqslant&\hat{P}_{t_k}\left(\left|\sum_{m=1}^{M_3} \gamma^{\tau - m} \epsilon_m S_{i,j}(t_k + m - 1) - \sum_{m=1}^{M_3} \gamma^{\tau - m} \epsilon_m E[S_{i,j}(t_k + \tau)]\right| \right. \\
&\left. > \sqrt{4 \hat{N}_{i,j}(t_k + \tau) U_{\mathrm{S}}^2 \log T} - \frac{U_{\mathrm{S}}}{8 T \log T}, \hat{N}_{i,j}(t_k + \tau) > 4 \log T\right)
\end{aligned}$$

$$\leqslant \hat{P}_{t_k}\Bigg(\left|\sum_{m=1}^{M_3}\gamma^{\tau-m}\epsilon_m\left(S_{i,j}(t_k+m-1)-E[S_{i,j}(t_k+m-1)]\right)\right| + \sum_{m=1}^{M_3}\gamma^{\tau-m}\epsilon_m\left|E[S_{i,j}(t_k+m-1)]-E[S_{i,j}(t_k+\tau)]\right| > \sqrt{4\hat{N}_{i,j}(t_k+\tau)U_{\mathrm{S}}^2\log T}-\frac{U_{\mathrm{S}}}{8T\log T}, \hat{N}_{i,j}(t_k+\tau)>4\log T\Bigg),$$

where in the last inequality we add and subtract the term $E[S_{i,j}(t_k+m-1)]$ and use the triangle inequality. We note that that in a stationary system, $E[S_{i,j}(t)]$ is a constant and the last step is not needed. Recall Assumption 1 on the time-varying service time. We have

$$\left|E[S_{i,j}(t_k+m-1)]-E[S_{i,j}(t_k+\tau)]\right| \leqslant \frac{1}{T}\left(\frac{1}{\gamma}\right)^{\tau-m}.$$

Hence, we have

$$\hat{P}_{t_k}\left(\mathcal{E}'^{\mathrm{c}}_{t_k,i,j,\tau}\right) \leqslant \hat{P}_{t_k}\Bigg(\left|\sum_{m=1}^{M_3}\gamma^{\tau-m}\epsilon_m\left(S_{i,j}(t_k+m-1)-E[S_{i,j}(t_k+m-1)]\right)\right| > \sqrt{4\hat{N}_{i,j}(t_k+\tau)U_{\mathrm{S}}^2\log T}-\frac{U_{\mathrm{S}}}{8T\log T}-\frac{M_3}{T}, \hat{N}_{i,j}(t_k+\tau)>4\log T\Bigg). \tag{66}$$

Recall that $t_k+M_3=f_j(t_k+\tau)$. Hence, $t_k+M_3\in[t_k+\tau-U_{\mathrm{S}}+1, t_k+\tau]$. Since $U_{\mathrm{S}}\leqslant\frac{T}{8}$ and $\tau\in[D_k-\frac{T}{8}, D_k+D_{k+1}-1]$, we have $D_k-\frac{T}{4}+1\leqslant M_3\leqslant D_k+D_{k+1}-1$. By the bound on D_k and D_{k+1} in Lemma 3, we further have

$$\frac{T}{4}+1\leqslant M_3\leqslant 2T-1. \tag{67}$$

Hence, we have

$$\frac{U_{\mathrm{S}}}{8T\log T}+\frac{M_3}{T}\leqslant\frac{U_{\mathrm{S}}}{8T\log T}+2\leqslant(2-\sqrt{3})\sqrt{4U_{\mathrm{S}}^2(\log T)^2}\leqslant(2-\sqrt{3})\sqrt{U_{\mathrm{S}}^2\hat{N}_{i,j}(t_k+\tau)\log T}, \tag{68}$$

where the second inequality holds for $T\geqslant e^5$, and the last inequality holds when $\hat{N}_{i,j}(t_k+\tau)>4\log T$. Based on (68), we can continue to bound (66) and obtain

$$\hat{P}_{t_k}\left(\mathcal{E}'^{\mathrm{c}}_{t_k,i,j,\tau}\right) \leqslant \hat{P}_{t_k}\left(\left|\sum_{m=1}^{M_3}\gamma^{\tau-m}\epsilon_m\left(S_{i,j}(t_k+m-1)-E[S_{i,j}(t_k+m-1)]\right)\right| > \sqrt{3\hat{N}_{i,j}(t_k+\tau)U_{\mathrm{S}}^2\log T}\right). \tag{69}$$

Note that $\hat{N}_{i,j}(t_k+\tau)\geqslant\sum_{m=1}^{M_3}\gamma^{\tau-m}\mathbb{1}_{\hat{i}_j^*(t_k+m-1)=i}\,\eta_j(t_k+m-1)=\sum_{m=1}^{M_3}\gamma^{\tau-m}\epsilon_m$. Hence, we can further bound (69) as

$$\hat{P}_{t_k}\left(\mathcal{E}'^{\mathrm{c}}_{t_k,i,j,\tau}\right) \leqslant \hat{P}_{t_k}\left(\frac{\left|\sum_{m=1}^{M_3}\gamma^{\tau-m}\epsilon_m\left(S_{i,j}(t_k+m-1)-E[S_{i,j}(t_k+m-1)]\right)\right|}{\sqrt{\sum_{m=1}^{M_3}\gamma^{\tau-m}\epsilon_m}} > \sqrt{3U_{\mathrm{S}}^2\log T}\right).$$

Since $M_3\leqslant\tau$ and $\gamma<1$, we have $\sqrt{\gamma^{M_3-\tau}}\geqslant 1$. Hence, we have

$$\begin{aligned}\hat{P}_{t_k}\left(\mathcal{E}'^{\mathrm{c}}_{t_k,i,j,\tau}\right) &\leqslant \hat{P}_{t_k}\left(\frac{\gamma^{M_3-\tau}\left|\sum_{m=1}^{M_3}\gamma^{\tau-m}\epsilon_m\left(S_{i,j}(t_k+m-1)-E[S_{i,j}(t_k+m-1)]\right)\right|}{\sqrt{\gamma^{M_3-\tau}}\sqrt{\sum_{m=1}^{M_3}\gamma^{\tau-m}\epsilon_m}} > \sqrt{3U_{\mathrm{S}}^2\log T}\right)\\ &= \hat{P}_{t_k}\left(\frac{\left|\sum_{m=1}^{M_3}\gamma^{M_3-m}\epsilon_m\left(S_{i,j}(t_k+m-1)-E[S_{i,j}(t_k+m-1)]\right)\right|}{\sqrt{\sum_{m=1}^{M_3}\gamma^{M_3-m}\epsilon_m}} > \sqrt{3U_{\mathrm{S}}^2\log T}\right)\end{aligned}$$

$$\leqslant \hat{P}_{t_k}\left(\frac{\left|\sum_{m=1}^{M_3}\gamma^{M_3-m}\epsilon_m\left(S_{i,j}(t_k+m-1)-E[S_{i,j}(t_k+m-1)]\right)\right|}{\sqrt{\sum_{m=1}^{M_3}\gamma^{2(M_3-m)}\epsilon_m}} > \sqrt{3U_{\mathrm{S}}^2\log T}\right), \tag{70}$$

where we added "2" in the last inequality because we want to use the Hoeffding-type inequality for self-normalized means (Garivier and Moulines, 2008, Theorem 18) later in the proof.

Consider the event $\mathcal{E}'_{t_k,i,j} \coloneqq \bigcap_{\tau=D_k-\frac{T}{8}}^{D_k+D_{k+1}-1}\mathcal{E}'_{t_k,i,j,\tau}$. Then from the result (70), we have

$$\begin{aligned}
&\hat{P}_{t_k}\left(\mathcal{E}'^{\mathrm{c}}_{t_k,i,j}\right)\\
\leqslant&\hat{P}_{t_k}\Bigg(\text{there exists } \tau\in\left[D_k-\frac{T}{8},D_k+D_{k+1}-1\right],\\
&\frac{\left|\sum_{m=1}^{M_3}\gamma^{M_3-m}\epsilon_m\left(S_{i,j}(t_k+m-1)-E[S_{i,j}(t_k+m-1)]\right)\right|}{\sqrt{\sum_{m=1}^{M_3}\gamma^{2(M_3-m)}\epsilon_m}} > \sqrt{3U_{\mathrm{S}}^2\log T}\Bigg)\\
\leqslant&\hat{P}_{t_k}\Bigg(\text{there exists } M\in\left[\frac{T}{4}+1,2T-1\right],\\
&\frac{\left|\sum_{m=1}^{M}\gamma^{M-m}\epsilon_m\left(S_{i,j}(t_k+m-1)-E[S_{i,j}(t_k+m-1)]\right)\right|}{\sqrt{\sum_{m=1}^{M}\gamma^{2(M-m)}\epsilon_m}} > \sqrt{3U_{\mathrm{S}}^2\log T}\Bigg)\\
\leqslant&\sum_{M=\frac{T}{4}+1}^{2T-1}\hat{P}_{t_k}\left(\frac{\left|\sum_{m=1}^{M}\gamma^{M-m}\epsilon_m\left(S_{i,j}(t_k+m-1)-E[S_{i,j}(t_k+m-1)]\right)\right|}{\sqrt{\sum_{m=1}^{M}\gamma^{2(M-m)}\epsilon_m}} > \sqrt{3U_{\mathrm{S}}^2\log T}\right).
\end{aligned} \tag{71}$$

where the second inequality uses the bound (67) on M_3, and the last inequality uses the union bound.

Let us view the conditional probability $\hat{P}_{t_k}$ as a new probability measure. Then $\hat{E}_{t_k}$ is the expectation under this measure. Note that $(S_{i,j}(t_k+m-1))_{m=1}^{\infty}$ is a sequence of independent bounded random variables under this new measure since they are independent of $\boldsymbol{Q}(t_k)$ and $\boldsymbol{H}(t_k)$, which also implies that

$$E[S_{i,j}(t_k+m-1)] = \hat{E}_{t_k}[S_{i,j}(t_k+m-1)]. \tag{72}$$

Let $\mathcal{F}_m$ defined as

$$\mathcal{F}_m \coloneqq \sigma\left((\boldsymbol{S}(t_k+n-1))_{n=1}^{m},(\boldsymbol{A}(t_k+n-1))_{n=1}^{m+1},(\boldsymbol{Q}(t_k+n-1))_{n=1}^{m+1},(\boldsymbol{H}(t_k+n-1))_{n=1}^{m+1}\right)$$

where $\sigma(\cdot)$ denotes the σ-algebra generated by the random variables. Note that

$$\sigma(S_{i,j}(t_k),...,S_{i,j}(t_k+m-1))\subset\mathcal{F}_m$$

and for any $n>m$, $S_{i,j}(t_k+n-1)$ is independent of $\mathcal{F}_m$. Recall that $\epsilon_m \coloneqq \mathbb{1}_{\hat{i}_j^*(t_k+m-1)=i}\eta_j(t_k+m-1)$. Since the scheduling decision at time t_k+m-1 is determined by $\boldsymbol{Q}(t_k+m-1)$ and $\boldsymbol{H}(t_k+m-1)$, $\mathbb{1}_{\hat{i}_j^*(t_k+m-1)=i}$ is $\mathcal{F}_{m-1}$-measurable. Since $\eta_j(t_k+m-1)$ is determined by $\boldsymbol{A}(t_k+m-1)$, $\boldsymbol{Q}(t_k+m-1)$, and $\boldsymbol{H}(t_k+m-1)$, $\eta_j(t_k+m-1)$ is also $\mathcal{F}_{m-1}$-measurable. Therefore, ϵ_m is $\mathcal{F}_{m-1}$-measurable, i.e., $(\epsilon_m)_{m=1}^{\infty}$ is a previsible (or predictable) sequence of Bernoulli random variables. We restate the Hoeffding-type inequality for self-normalized means in (Garivier and Moulines, 2008, Theorem 18), (Garivier and Moulines, 2011) in our setting as follows:

Theorem 2 (Hoeffding-type inequality for self-normalized means, Theorem 18 in (Garivier and Moulines, 2008))**.** *Let $(X_m)_{m\geqslant 1}$ be a sequence of nonnegative independent bounded random variables with $X_m\in[0,B]$. $\sigma(X_1,\ldots,X_m)\subset\mathcal{F}_m$ and for $n>m$, X_n is independent of $\mathcal{F}_m$. For all integers M and all $\beta>0$,*

$$\hat{P}_{t_k}\left(\frac{\sum_{m=1}^{M}\gamma^{M-m}X_m\epsilon_m-\sum_{m=1}^{M}\gamma^{M-m}\hat{E}_{t_k}[X_m]\epsilon_m}{\sqrt{\sum_{m=1}^{M}\gamma^{2(M-m)}\epsilon_m}}>\beta\right)$$

$$\leqslant \left\lceil \frac{\log \sum_{m=1}^{M} \gamma^{M-m}}{\log(1+\zeta)} \right\rceil \exp\left(-\frac{2\beta^2}{B^2}\left(1-\frac{\zeta^2}{16}\right)\right)$$

for all $\zeta > 0$.

Applying Theorem 2 with $X_m = S_{i,j}(t_k+m-1)$, $\beta = \sqrt{3U_{\mathrm{S}}^2 \log T}$, and $B = U_{\mathrm{S}}$, we have

$$\begin{aligned}
&\hat{P}_{t_k}\left(\frac{\sum_{m=1}^{M}\gamma^{M-m}\epsilon_m\left(S_{i,j}(t_k+m-1)-\hat{E}_{t_k}[S_{i,j}(t_k+m-1)]\right)}{\sqrt{\sum_{m=1}^{M}\gamma^{2(M-m)}\epsilon_m}} > \sqrt{3U_{\mathrm{S}}^2\log T}\right)\\
\leqslant &\left(\frac{\log\sum_{m=1}^{M}\gamma^{M-m}}{\log(1+\zeta)}+1\right)\exp\left(-6\left(1-\frac{\zeta^2}{16}\right)\log T\right)\\
\leqslant &\left(\frac{\log(\frac{T}{8\log T})}{\log(1+\zeta)}+1\right)\exp\left(-6\left(1-\frac{\zeta^2}{16}\right)\log T\right)
\end{aligned}$$

for all $\zeta > 0$ and all positive integers M, where the last inequality holds since $\sum_{m=1}^{M}\gamma^{M-m} = \frac{1-\gamma^M}{1-\gamma} \leqslant \frac{1}{1-\gamma} = \frac{T}{8\log T}$. Although in (Garivier and Moulines, 2008) the bound is only proved for overestimation, the proof can be extended to show that the bound also holds for underestimation. Specifically, note that

$$\hat{E}_{t_k}\left[\hat{E}_{t_k}[S_{i,j}(t_k+m-1)] - S_{i,j}(t_k+m-1)\right] = 0$$

and

$$\hat{E}_{t_k}[S_{i,j}(t_k+m-1)] - U_{\mathrm{S}} \leqslant \hat{E}_{t_k}[S_{i,j}(t_k+m-1)] - S_{i,j}(t_k+m-1) \leqslant \hat{E}_{t_k}[S_{i,j}(t_k+m-1)].$$

Hence, considering the random variable $\hat{E}_{t_k}[S_{i,j}(t_k+m-1)] - S_{i,j}(t_k+m-1)$, from (Devroye et al., 1996, Lemma 8.1), for any $\lambda > 0$, we have

$$\log \hat{E}_{t_k}[\exp(-\lambda S_{i,j}(t_k+m-1))] \leqslant \frac{\lambda^2 U_{\mathrm{S}}^2}{8} - \lambda \hat{E}_{t_k}[S_{i,j}(t_k+m-1)].$$

Hence, we can apply the same proof in (Garivier and Moulines, 2008, Theorem 18) by replacing λ in the proof with $-\lambda$. Then for underestimation, we also have the same bound, i.e.,

$$\begin{aligned}
&\hat{P}_{t_k}\left(\frac{\sum_{m=1}^{M}\gamma^{M-m}\epsilon_m\left(\hat{E}_{t_k}[S_{i,j}(t_k+m-1)]-S_{i,j}(t_k+m-1)\right)}{\sqrt{\sum_{m=1}^{M}\gamma^{2(M-m)}\epsilon_m}} > \sqrt{3U_{\mathrm{S}}^2\log T}\right)\\
\leqslant &\left(\frac{\log(\frac{T}{8\log T})}{\log(1+\zeta)}+1\right)\exp\left(-6\left(1-\frac{\zeta^2}{16}\right)\log T\right)
\end{aligned}$$

for all $\zeta > 0$ and all positive integers M. Taking the union bound over underestimation and overestimation, we have

$$\begin{aligned}
&\hat{P}_{t_k}\left(\frac{\left|\sum_{m=1}^{M}\gamma^{M-m}\epsilon_m\left(S_{i,j}(t_k+m-1)-\hat{E}_{t_k}[S_{i,j}(t_k+m-1)]\right)\right|}{\sqrt{\sum_{m=1}^{M}\gamma^{2(M-m)}\epsilon_m}} > \sqrt{3U_{\mathrm{S}}^2\log T}\right)\\
\leqslant &2\left(\frac{\log(\frac{T}{8\log T})}{\log(1+\zeta)}+1\right)\exp\left(-6\left(1-\frac{\zeta^2}{16}\right)\log T\right)
\end{aligned}$$

for all $\zeta > 0$ and all positive integers M. Setting $\zeta = 0.3$, we have

$$\hat{P}_{t_k}\left(\frac{\left|\sum_{m=1}^{M}\gamma^{M-m}\epsilon_m\left(S_{i,j}(t_k+m-1)-\hat{E}_{t_k}[S_{i,j}(t_k+m-1)]\right)\right|}{\sqrt{\sum_{m=1}^{M}\gamma^{2(M-m)}\epsilon_m}} > \sqrt{3U_{\mathrm{S}}^2\log T}\right) \leqslant \frac{2\log T}{T^5\log 1.3} \tag{73}$$

for all $T \geqslant e^5$ and all positive integers M.

Combining (71), (72), and (73), we have

$$\hat{P}_{t_k}\left(\mathcal{E}'^{\mathrm{c}}_{t_k,i,j}\right) \leqslant \sum_{M=\frac{T}{4}+1}^{2T-1} \frac{2\log T}{T^5 \log 1.3} \leqslant \frac{1}{T^2}$$

for all $T \geqslant e^5$. Taking the union bound over i, we have

$$\hat{P}_{t_k}(\mathcal{E}'^{\mathrm{c}}_{t_k,j}) \leqslant \sum_{i=1}^{I} \hat{P}_{t_k}\left(\mathcal{E}'^{\mathrm{c}}_{t_k,i,j}\right) \leqslant \frac{I}{T^2}.$$

□

C.2 Proof of Lemma 2

Lemma 2. *For any j and any $k \geqslant 0$,*

$$\sum_{\tau=D_k}^{D_k+D_{k+1}-1} b_{I_j(t_k+\tau),j}(f_j(t_k+\tau))\eta_j(f_j(t_k+\tau))\mathbb{1}_{I_j(t_k+\tau),j}(t_k+\tau) \leqslant 99IU_{\mathrm{S}}\sqrt{T}\log T.$$

Proof. Notice that the summation can be rewritten as

$$\begin{aligned}
&\sum_{\tau=D_k}^{D_k+D_{k+1}-1} b_{I_j(t_k+\tau),j}(f_j(t_k+\tau))\eta_j(f_j(t_k+\tau))\mathbb{1}_{I_j(t_k+\tau),j}(t_k+\tau) \\
=&\sum_i \sum_{\tau=D_k}^{D_k+D_{k+1}-1} b_{i,j}(f_j(t_k+\tau))\eta_j(f_j(t_k+\tau))\mathbb{1}_{i,j}(t_k+\tau)\mathbb{1}_{I_j(t_k+\tau)=i}. \qquad (74)
\end{aligned}$$

The summation term for queue i can be rewritten as

$$\sum_{\tau=D_k}^{D_k+D_{k+1}-1} b_{i,j}(f_j(t_k+\tau))\eta_j(f_j(t_k+\tau))\mathbb{1}_{i,j}(t_k+\tau)\mathbb{1}_{I_j(t_k+\tau)=i} = \sum_{n=1}^{N} b_{i,j}(\tau_n), \qquad (75)$$

where N is the total number of times when server j picks queue i such that server j is not idling and the completion time is in the time interval $[t_k + D_k, t_k + D_k + D_{k+1} - 1]$, and τ_n is the time slot for the n^{th} time.

From (33), we have $f_j(t_k+\tau) \in [t_k, t_k+2T-1]$, and thus $\tau_n \in [t_k, t_k+2T-1]$. Divide the interval $[t_k, t_k+2T-1]$ into parts where each part contains $\left\lceil \frac{T}{16\log T} \right\rceil$ samples. Then there are at most $\lceil 32\log T \rceil$ parts. Consider the m^{th} part such that

$$\tau_n \in \left[t_k + (m-1)\left\lceil \frac{T}{16\log T} \right\rceil, t_k + m\left\lceil \frac{T}{16\log T} \right\rceil - 1 \right].$$

Consider the summation in the m^{th} part:

$$\sum_{n=1}^{N_m} b_{i,j}(\tau_{m,n}),$$

where N_m is the total number of times in the m^{th} part such that $\sum_m N_m = N$, and $\tau_{m,n}$ is the time slot for the n^{th} time in the m^{th} part. If $N_m = 0$, then $\sum_{n=1}^{N_m} b_{i,j}(\tau_{m,n}) = 0$. We now consider the case where $N_m > 0$ and derive an upper bound for the summation. First, we know that

$$\hat{N}_{i,j}(\tau_{m,1}) \geqslant 0, \text{ and } b_{i,j}(\tau_{m,1}) \leqslant 1.$$

Consider the contribution of the completion of the job starting at $\tau_{m,1}$ to $\hat{N}_{i,j}(\tau_{m,2})$. Since $\tau_{m,2}-\tau_{m,1} \leqslant \left\lceil \frac{T}{16\log T} \right\rceil$ and $\frac{T}{16\log T} \geqslant 1$, we have

$$\hat{N}_{i,j}(\tau_{m,2}) \geqslant \left(1-\frac{8\log T}{T}\right)^{\tau_{m,2}-\tau_{m,1}-1} \geqslant \frac{1}{2},$$

where the last inequality follows from the fact that $(1-x)^y \geqslant 1-xy$ for any $x \in [0,1]$ and $y \geqslant 1$. For a general n, we have

$$\hat{N}_{i,j}(\tau_{m,n+1}) \geqslant \sum_{l=1}^{n} \left(1-\frac{8\log T}{T}\right)^{\tau_{m,n+1}-\tau_{m,l}-1}.$$

Since $\tau_{m,n+1}-\tau_{m,l} \leqslant \left\lceil \frac{T}{16\log T} \right\rceil$ for any l, we have

$$\left(1-\frac{8\log T}{T}\right)^{\tau_{m,n+1}-\tau_{m,l}-1} \geqslant \frac{1}{2},$$

and thus

$$\hat{N}_{i,j}(\tau_{m,n+1}) \geqslant \sum_{l=1}^{n} \left(1-\frac{8\log T}{T}\right)^{\tau_{m,n+1}-\tau_{m,l}-1} \geqslant \frac{n}{2}.$$

Hence, we have

$$b_{i,j}(\tau_{m,n+1}) \leqslant \sqrt{\frac{16U_{\mathrm{S}}^2\log T}{\hat{N}_{i,j}(\tau_{m,n+1})}} \leqslant U_{\mathrm{S}}\sqrt{\frac{32\log T}{n}}.$$

From the analysis above, we obtain

$$\sum_{n=1}^{N_m} b_{i,j}(\tau_{m,n}) \leqslant 1+\sum_{n=2}^{N_m} U_{\mathrm{S}}\sqrt{\frac{32\log T}{n-1}} \leqslant 1+U_{\mathrm{S}}\sqrt{128(N_m-1)\log T} \leqslant 1+\sqrt{8T}U_{\mathrm{S}}$$

where the last inequality follows from the fact that $N_m \leqslant \left\lceil \frac{T}{16\log T} \right\rceil$.

Therefore, since there are at most $\lceil 32\log T \rceil$ parts, we conclude that

$$\sum_{n=1}^{N} b_{i,j}(\tau_n) \leqslant \lceil 32\log T \rceil \left(1+\sqrt{8T}U_{\mathrm{S}}\right) \leqslant 99U_{\mathrm{S}}\sqrt{T}\log T.$$

Combining the above bound with (74) and (75), we have

$$\sum_{\tau=D_k}^{D_k+D_{k+1}-1} b_{I_j(t_k+\tau),j}(f_j(t_k+\tau))\eta_j(f_j(t_k+\tau))\mathbb{1}_{I_j(t_k+\tau),j}(t_k+\tau) \leqslant 99IU_{\mathrm{S}}\sqrt{T}\log T.$$

□

C.3 Proof of Lemma 3

Lemma 3. *Suppose $W \leqslant \frac{T}{2}$. Then $\frac{T}{2} \leqslant D(t) \leqslant \frac{T}{2}+W \leqslant T$ for any t.*

Proof. Recall the definition of $D(t)$:

$$D(t) = \min_n \sum_{l=0}^{n} w(\tau_l(t))$$

$$\text{s.t. } \sum_{l=0}^{n} w(\tau_l(t)) \geqslant \frac{T}{2}.$$

Recall that $n^*(t)$ is the optimal solution to the above optimization problem. Note that $D(t) = \sum_{l=0}^{n^*(t)} w(\tau_l(t)) \geqslant \frac{T}{2}$ and $\sum_{l=0}^{n^*(t)-1} w(\tau_l(t)) < \frac{T}{2}$. Hence, we have

$$D(t) = \sum_{l=0}^{n^*(t)} w(\tau_l(t)) = \sum_{l=0}^{n^*(t)-1} w(\tau_l(t)) + w(\tau_{n^*(t)}(t)) \leqslant \frac{T}{2} + W,$$

where the last inequality is due to the bound $w(\tau) \leqslant W$ for any τ. Therefore, for any t, we have

$$\frac{T}{2} \leqslant D(t) \leqslant \frac{T}{2} + W \leqslant T,$$

where the last inequality is by $W \leqslant \frac{T}{2}$. □

C.4 Proof of Lemma 4

Lemma 4. *For any i, t, $Q_i(t+1) \leqslant \max\left\{J, Q_i(t) + A_i(t) - \sum_j \mathbb{1}_{i,j}(t)\right\}$.*

Proof. Fix i and t. Consider two cases. The first case is that there exists j such that $\eta_j(t) = 0$ (server j is idling) and $I_j(t) = i$. The second case is that for all servers j, $\eta_j(t) = 1$ or $I_j(t) \neq i$.

Notice that for the first case we must have

$$\tilde{Q}_i(t) + A_i(t) = 0$$

since server j is scheduled to i and is idling. Hence, we have

$$Q_i(t) + A_i(t) \leqslant \tilde{Q}_i(t) + J + A_i(t) = J.$$

Hence, by the queue dynamics (1) and the above inequality, we have

$$Q_i(t+1) \leqslant Q_i(t) + A_i(t) \leqslant J$$

for the first case. For the second case, we have

$$\begin{aligned} Q_i(t+1) =& Q_i(t) + A_i(t) - \sum_j \mathbb{1}_{i,j}(t)\eta_j(t) \\ =& Q_i(t) + A_i(t) - \sum_j \mathbb{1}_{i,j}(t), \end{aligned}$$

where the second inequality holds since for any server j, either $\eta_j(t) = 1$ or $\mathbb{1}_{i,j}(t) = 0$.

Combining the two cases, we obtain that for any i, t,

$$Q_i(t+1) \leqslant \max\left\{J, Q_i(t) + A_i(t) - \sum_j \mathbb{1}_{i,j}(t)\right\}.$$

□

C.5 Proof of Lemma 5

Lemma 5. *For any t, i, $\tau \geqslant 0$, we have*

1. *$Q_i(t) - J\tau \leqslant Q_i(t+\tau) \leqslant Q_i(t) + \tau U_{\mathrm{A}}$;*
2. *$\sum_i Q_i(t+\tau) \geqslant \sum_i Q_i(t) - J\tau$.*

Proof. (1) holds since $Q_i(t)$ can increase at most U_{A} and can decrease at most J for each time slot by the queue dynamics (1). (2) holds since the total queue length can decrease by at most J for each time slot. This is because there are J servers in total and each server can serve at most one job at a time.

□

C.6 Proof of Lemma 6

Lemma 6. $\sum_i q_i \leqslant \frac{1}{D_{k+1}} \sum_{\tau=D_k}^{D_k+D_{k+1}-1} \hat{E}_{t_k} \left[\sum_i Q_i(t_k+\tau) + 2JT\right]$.

Proof. By Lemma 5, we have

$$\sum_i Q_i(t_k+\tau) \geqslant \sum_i Q_i(t_k) - J\tau \geqslant \sum_i Q_i(t_k) - 2JT, \tag{76}$$

where the last inequality holds since $\tau \leqslant D_k + D_{k+1} - 1 \leqslant 2T$ by Lemma 3. Based on (76), we have

$$\sum_i q_i \leqslant \frac{1}{D_{k+1}} \sum_{\tau=D_k}^{D_k+D_{k+1}-1} \hat{E}_{t_k} \left[\sum_i Q_i(t_k+\tau) + 2JT\right].$$

□

C.7 Proof of Lemma 7

Lemma 7. *For any $k \geqslant 0$, $i \in \{1,\dots,I\}$, $j \in \{1,\dots,J\}$, we have*

$$\hat{P}_{t_k}\left(\mathcal{E}^{\mathrm{c}}_{X,t_k,i,j}\right) \coloneqq \Pr\left(\mathcal{E}^{\mathrm{c}}_{X,t_k,i,j} \,|\mathbf{Q}(t_k)=\mathbf{q}, \mathbf{H}(t_k)=\mathbf{h}\right) \leqslant \frac{1}{T^2},$$

where

$$\mathcal{E}_{X,t_k,i,j} \coloneqq \left\{ \sum_{\tau=D_k}^{D_k+D_{k+1}-1} v_j(t_k+\tau) \mathbb{1}_{\hat{i}^*_j(t_k+\tau)=i} \eta_j(t_k+\tau) \left(E[X_{i,j}(t_k+\tau)] - X_{i,j}(t_k+\tau)\right) \geqslant -U_{\mathrm{S}}\sqrt{2T\log T}\sum_i q_i \right\}.$$

Proof. If server j is not idling, then $X_{i,j}(t) = S_{i,j}(t)$ by definition. Also note that $X_{i,j}(t) \leqslant S_{i,j}(t)$ by definition. Hence, we have

$$\begin{aligned}
&\mathcal{E}_{X,t_k,i,j} \\
\subseteq &\left\{ \sum_{\tau=D_k}^{D_k+D_{k+1}-1} v_j(t_k+\tau) \mathbb{1}_{\hat{i}^*_j(t_k+\tau)=i} \eta_j(t_k+\tau) \left(E[S_{i,j}(t_k+\tau)] - S_{i,j}(t_k+\tau)\right) \geqslant -U_{\mathrm{S}}\sqrt{2T\log T}\sum_i q_i \right\} \\
= &\Bigg\{ \sum_{m=1}^{D_{k+1}} v_j(t_k+D_k+m-1) \mathbb{1}_{\hat{i}^*_j(t_k+D_k+m-1)=i} \eta_j(t_k+D_k+m-1) \\
&\qquad \left(E[S_{i,j}(t_k+D_k+m-1)] - S_{i,j}(t_k+D_k+m-1)\right) \geqslant -U_{\mathrm{S}}\sqrt{2T\log T}\sum_i q_i \Bigg\}.
\end{aligned}$$

Let $\epsilon_m \coloneqq \mathbb{1}_{\hat{i}^*_j(t_k+D_k+m-1)=i}\eta_j(t_k+D_k+m-1)$. Let $S^{(m)}_{i,j} \coloneqq S_{i,j}(t_k+D_k+m-1)$. Let $v^{(m)}_j \coloneqq v_j(t_k+D_k+m-1)$. Then

$$\begin{aligned}
\hat{P}_{t_k}\left(\mathcal{E}^{\mathrm{c}}_{X,t_k,i,j}\right) \leqslant &\hat{P}_{t_k}\left(\sum_{m=1}^{D_{k+1}} \epsilon_m v^{(m)}_j \left(E[S^{(m)}_{i,j}] - S^{(m)}_{i,j}\right) < -U_{\mathrm{S}}\sqrt{2T\log T}\sum_i q_i\right) \\
= &\hat{P}_{t_k}\left(\frac{\sum_{m=1}^{D_{k+1}} \epsilon_m v^{(m)}_j \left(E[S^{(m)}_{i,j}] - S^{(m)}_{i,j}\right)}{\sqrt{T}} < -U_{\mathrm{S}}\sqrt{2\log T}\sum_i q_i\right)
\end{aligned}$$

$$\leqslant \hat{P}_{t_k}\left(\frac{\sum_{m=1}^{D_{k+1}}\epsilon_m v_j^{(m)}\left(E[S_{i,j}^{(m)}]-S_{i,j}^{(m)}\right)}{\sqrt{\sum_{m=1}^{D_{k+1}}\epsilon_m}} < -U_{\mathrm{S}}\sqrt{2\log T}\sum_i q_i\right)$$
$$= \hat{P}_{t_k}\left(\frac{\sum_{m=1}^{D_{k+1}}\epsilon_m v_j^{(m)}\left(S_{i,j}^{(m)}-E[S_{i,j}^{(m)}]\right)}{\sqrt{\sum_{m=1}^{D_{k+1}}\epsilon_m}} > U_{\mathrm{S}}\sqrt{2\log T}\sum_i q_i\right), \tag{77}$$

where the last inequality holds since $\sum_{m=1}^{D_{k+1}}\epsilon_m \leqslant D_{k+1} \leqslant T$ by Lemma 3.

Let us view the conditional probability $\hat{P}_{t_k}$ as a new probability measure. Then $\hat{E}_{t_k}$ is the expectation under this measure. Note that $\left(S_{i,j}^{(m)}\right)_{m=1}^{\infty}$ is a sequence of independent bounded random variables under this new measure since they are independent of $\boldsymbol{Q}(t_k)$ and $\boldsymbol{H}(t_k)$, which also implies that

$$E\left[S_{i,j}^{(m)}\right] = \hat{E}_{t_k}\left[S_{i,j}^{(m)}\right].$$

Note that given $\boldsymbol{Q}(t_k)$ and $\boldsymbol{H}(t_k)$, $v_j^{(m)}$ is a constant, which implies

$$v_j^{(m)}\hat{E}_{t_k}\left[S_{i,j}^{(m)}\right] = \hat{E}_{t_k}\left[v_j^{(m)}S_{i,j}^{(m)}\right].$$

Therefore, from (77), we have

$$\hat{P}_{t_k}\left(\mathcal{E}^{\mathrm{c}}_{X,t_k,i,j}\right) \leqslant \hat{P}_{t_k}\left(\frac{\sum_{m=1}^{D_{k+1}}\epsilon_m\left(v_j^{(m)}S_{i,j}^{(m)}-\hat{E}_{t_k}[v_j^{(m)}S_{i,j}^{(m)}]\right)}{\sqrt{\sum_{m=1}^{D_{k+1}}\epsilon_m}} > U_{\mathrm{S}}\sqrt{2\log T}\sum_i q_i\right). \tag{78}$$

Let $\mathcal{F}_m$ defined as

$$\mathcal{F}_m := \sigma\Big((\boldsymbol{S}(t_k+D_k+n-1))_{n=1}^{m}, (\boldsymbol{A}(t_k+D_k+n-1))_{n=1}^{m+1},$$
$$(\boldsymbol{Q}(t_k+D_k+n-1))_{n=1}^{m+1}, (\boldsymbol{H}(t_k+D_k+n-1))_{n=1}^{m+1}\Big)$$

where $\sigma(\cdot)$ denotes the σ-algebra generated by the random variables. Note that

$$\sigma\left(v_j^{(1)}S_{i,j}^{(1)}, v_j^{(2)}S_{i,j}^{(2)}, \ldots, v_j^{(m)}S_{i,j}^{(m)}\right) \subset \mathcal{F}_m$$

and for any $n > m$, $v_j^{(n)}S_{i,j}^{(n)}$ is independent of $\mathcal{F}_m$. Also note that $(\epsilon_m)_{m=1}^{\infty}$ is a previsible (predictable) sequence of Bernoulli random variables, i.e., ϵ_m is $\mathcal{F}_{m-1}$-measurable. Note that $0 \leqslant v_j^{(n)}S_{i,j}^{(n)} \leqslant U_{\mathrm{S}}\sum_i q_i$. Based on the above conditions, we can then apply Theorem 2 (Hoeffding-type inequality for self-normalized means, Theorem 18 in (Garivier and Moulines, 2008)) with $X_m = v_j^{(m)}S_{i,j}^{(m)}$, $\beta = U_{\mathrm{S}}\sqrt{2\log T}\sum_i q_i$, and $B = U_{\mathrm{S}}\sum_i q_i$ to obtain

$$\hat{P}_{t_k}\left(\frac{\sum_{m=1}^{D_{k+1}}\epsilon_m\left(v_j^{(m)}S_{i,j}^{(m)}-\hat{E}_{t_k}[v_j^{(m)}S_{i,j}^{(m)}]\right)}{\sqrt{\sum_{m=1}^{D_{k+1}}\epsilon_m}} > U_{\mathrm{S}}\sqrt{2\log T}\sum_i q_i\right)$$
$$\leqslant \left(\frac{\log D_{k+1}}{\log(1+\zeta)}+1\right)\exp\left(-4\left(1-\frac{\zeta^2}{16}\right)\log T\right).$$

Setting $\zeta = 0.3$ and by the bound on D_{k+1} in Lemma 3, we have

$$\hat{P}_{t_k}\left(\frac{\sum_{m=1}^{D_{k+1}}\epsilon_m\left(v_j^{(m)}S_{i,j}^{(m)}-\hat{E}_{t_k}[v_j^{(m)}S_{i,j}^{(m)}]\right)}{\sqrt{\sum_{m=1}^{D_{k+1}}\epsilon_m}} > U_{\mathrm{S}}\sqrt{2\log T}\sum_i q_i\right) \leqslant \frac{1}{T^2} \tag{79}$$

for $T \geqslant e^5$. Substituting (79) into (78), the lemma is proved.

□

D ADDITIONAL DETAILS OF THE SIMULATIONS

In this section, we present more details of the simulations.

D.1 Settings

For the stationary setting, we set the arrival rates $\lambda_i = 0.75$ for all $i \in \{1, 2, \ldots, 10\}$. We set the service rates as follows:

$$\mu_{2k+1,2l+1} = 0.9, \quad \mu_{2k+1,2l+2} = 0.6, \quad \mu_{2k+2,2l+1} = 0.5, \quad \mu_{2k+2,2l+2} = 1.0,$$

for all $k, l \in \{0, 1, 2, 3, 4\}$.

For the nonstationary aperiodic setting, we set the arrival rates $\lambda_i(t) = 0.70$ for all $i \in \{1, 2, \ldots, 10\}$ and all t. We set the service rates as follows:

$$\begin{aligned}
\mu_{2k+1,2l+1}(t) =& (0.90000, 0.89999, \ldots, 0.60001),\\
\mu_{2k+1,2l+2}(t) =& (0, 60000, 0.60001, \ldots, 0.89999),\\
\mu_{2k+2,2l+1}(t) =& (0.50000, 0.50001, \ldots, 0.79999),\\
\mu_{2k+2,2l+2}(t) =& (1.00000, 0.99999, \ldots, 0.70001),
\end{aligned}$$

for all $k, l \in \{0, 1, 2, 3, 4\}$.

For the nonstationary periodic setting, we set the arrival rates

$$\lambda_i(t) = (0.750, 0.749, \ldots, 0.551, 0.550, 0.551, \ldots, 0.749, 0.700, \ldots),$$

where the period is 400, for all $i \in \{1, 2, \ldots, 10\}$. We set the service rates as follows:

$$\begin{aligned}
\mu_{2k+1,2l+1}(t) =& (0.900, 0.899, \ldots, 0.501, 0.500, 0.501, \ldots, 0.899, 0.900, \ldots), \text{ period} = 800,\\
\mu_{2k+1,2l+2}(t) =& (0.600, 0.601, \ldots, 0.999, 1.000, 0.999, \ldots, 0.601, 0.600, \ldots) \text{ period} = 800,\\
\mu_{2k+2,2l+1}(t) =& (0, 500, 0.501, \ldots, 0.899, 0.900, 0.899, \ldots, 0.501, 0.500, \ldots) \text{ period} = 800,\\
\mu_{2k+2,2l+2}(t) =& (1.000, 0.999, \ldots, 0.601, 0.600, 0.601, \ldots, 0.999, 1.000, \ldots) \text{ period} = 800,
\end{aligned}$$

for all $k, l \in \{0, 1, 2, 3, 4\}$.

For the second nonstationary periodic setting with a larger period, we set the arrival rates

$$\lambda_i(t) = (0.7500, 0.7499, \ldots, 0.5501, 0.5500, 0.5501, \ldots, 0.7499, 0.7500, \ldots),$$

where the period is 4000, for all $i \in \{1, 2, \ldots, 10\}$. We set the service rates as follows:

$$\begin{aligned}
\mu_{2k+1,2l+1}(t) &= (0.9000, 0.8999, \ldots, 0.5001, 0.5000, 0.5001, \ldots, 0.8999, 0.9000, \ldots), \text{ period} = 8000,\\
\mu_{2k+1,2l+2}(t) &= (0.6000, 0.6001, \ldots, 0.9999, 1.0000, 0.9999, \ldots, 0.6001, 0.6000, \ldots), \text{ period} = 8000,\\
\mu_{2k+2,2l+1}(t) &= (0, 5000, 0.5001, \ldots, 0.8999, 0.9000, 0.8999, \ldots, 0.5001, 0.5000, \ldots), \text{ period} = 8000,\\
\mu_{2k+2,2l+2}(t) &= (1.0000, 0.9999, \ldots, 0.6001, 0.6000, 0.6001, \ldots, 0.9999, 1.0000, \ldots), \text{ period} = 8000,
\end{aligned}$$

for all $k, l \in \{0, 1, 2, 3, 4\}$.

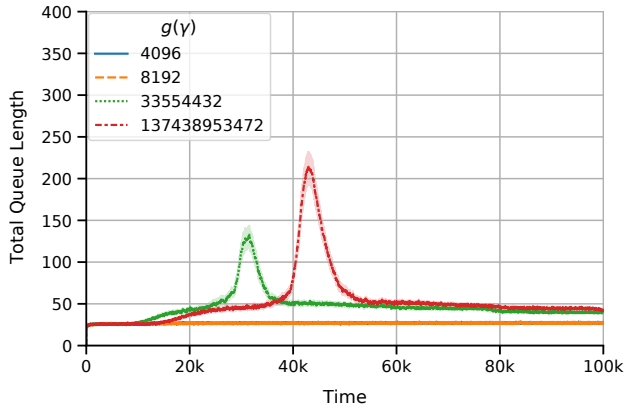
D.2 Parameters

For the stationary setting, we use the parameters that have stability guarantee in theory for both *MaxWeight with discounted UCB* and *DAM.UCB*. Specifically, for *MaxWeight with discounted UCB*, by Theorem 1, we choose $c_1 = 4$ and $g(\gamma) = 6071209191677812$ such that $g(\gamma)$ satisfies $\delta_{\max} \geqslant \frac{804IJU_S^2 \log g(\gamma)}{[g(\gamma)]^{1/2}}$, where $\delta_{\max} = 0.15$ is the largest δ for the arrivals and capacity region in our stationary setting. For *DAM.UCB*, the traffic slackness ϵ defined in (Freund et al., 2022) is 0.2 in our stationary setting, so according to Freund et al. (2022), the epoch size we choose is $L_{\text{epoch}} = \lceil (32/\epsilon + 1) L_{\text{conv}} \rceil = 161$, where we set $L_{\text{conv}} = 1$ since there is no need to use *DAM.converge* algorithm in their paper because we consider centralized setting. For *frame-based MaxWeight* (Stahlbuhk et al., 2019), since in their paper there is no theoretical value for the

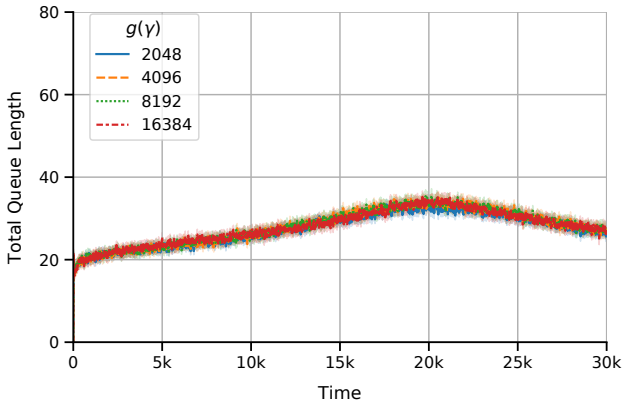
parameter frame size, we try different frame sizes and then choose the one that has the best performance, which is 120. For *MaxWeight with discounted EM*, we use the same γ as *MaxWeight with discounted UCB*.

For the nonstationary settings, we choose $c_1 = 4$ and $g(\gamma) = 8192$ for *MaxWeight with discounted UCB*. For *DAM.UCB*, we use $L_{\text{epoch}} = 161$ for the periodic setting and $L_{\text{epoch}} = 113$ in the aperiodic setting, which are both calculated from $L_{\text{epoch}} = \lceil (32/\epsilon + 1) L_{\text{conv}} \rceil$ according to (Freund et al., 2022) using the corresponding traffic slackness at $t = 0$. For *frame-based MaxWeight* (Stahlbuhk et al., 2019), same as in the stationary setting, we try different frame sizes and then choose the one that has the best performance, which are 30, 25, 20 for the settings of aperiodic, periodic, and periodic with larger period, respectively. For *MaxWeight with discounted EM*, we use the same γ as *MaxWeight with discounted UCB*.

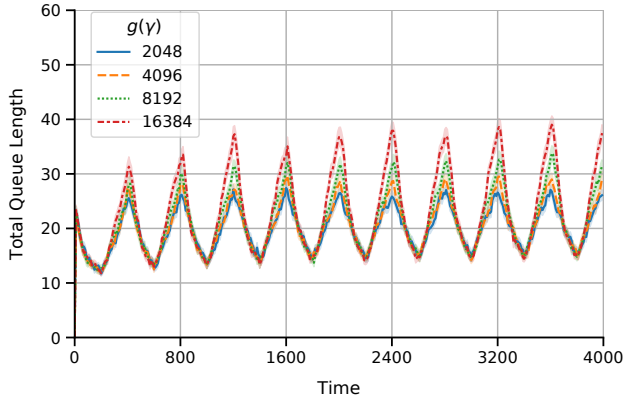
Simulation results of *MaxWeight with discounted UCB* with different $g(\gamma)$ are shown in Fig. 5. As shown in the figures, the proposed algorithm works well under different values of $g(\gamma)$ (hence, γ). Therefore, the algorithm is robust to the value of γ. Note that these values are chosen somewhat arbitrarily, not optimized. They are 2 to the power of some arbitrary integer.



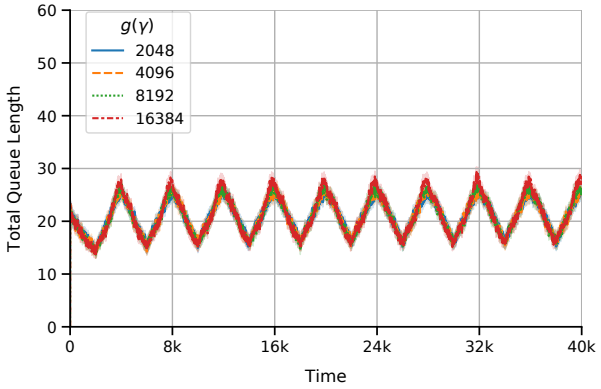
(a) Stationary.



(b) Nonstationary Aperiodic.



(c) Nonstationary Periodic.



(d) Nonstationary Periodic.

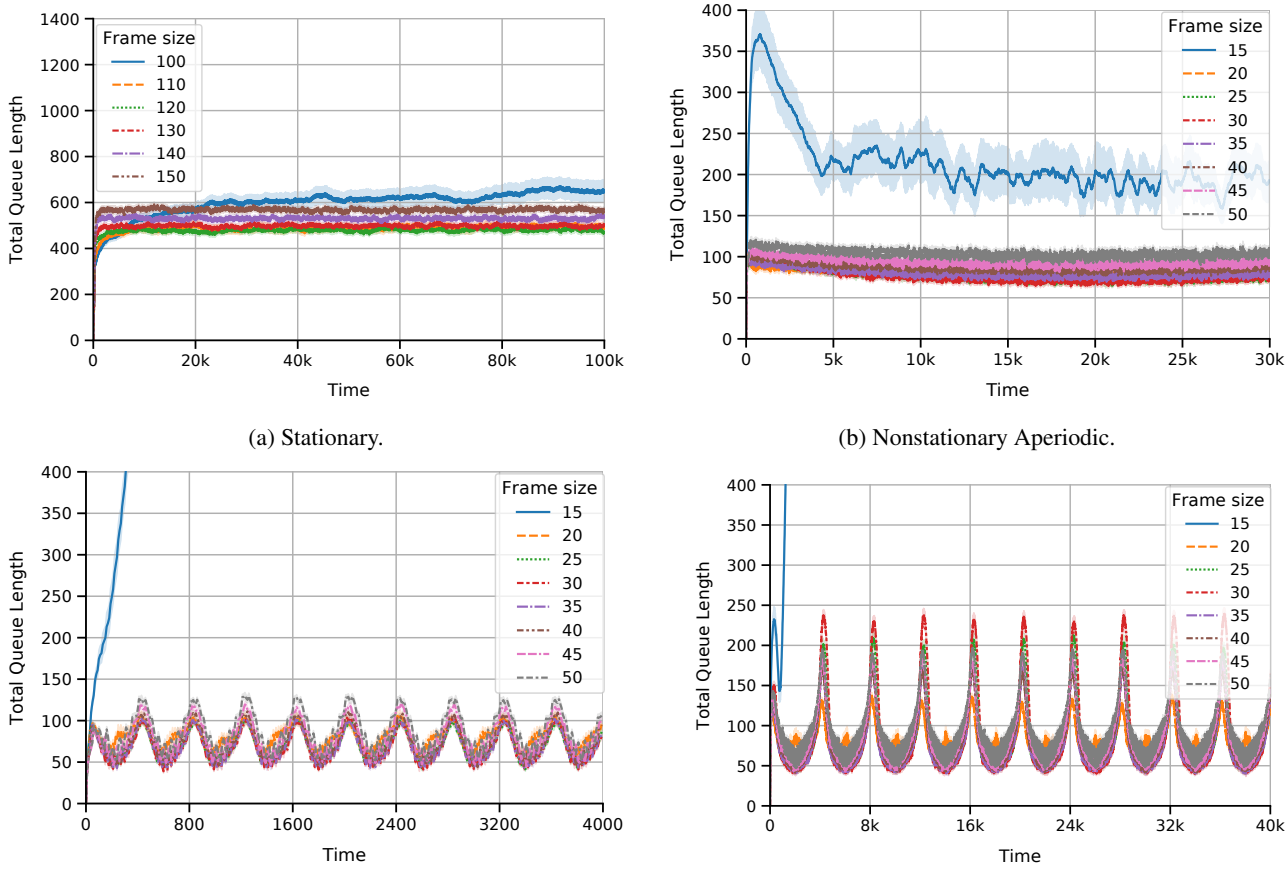
Figure 5: Simulation Results of *MaxWeigh With Discounted UCB* With Different $g(\gamma)$.

In order to choose the best frame size parameter for the *frame-based MaxWeight* algorithm, we conducted simulations of *frame-based MaxWeight* algorithm with different frame sizes. The results are shown in Fig. 6.

The total queue length $\sum_i Q_i(t)$ of all the curves in the figures is averaged over 100 runs. The shaded area in all the figures is the 95% confidence interval. For all the curves, we plot one point every 10 time slots.

D.3 Additional Figures

We present all the comparison results in Fig. 7, Fig 8, Fig. 9, and Fig, 10, each with a zoom in view and a zoom out view. The zoom out view has a Y-axis with a larger range so that it can include the parts of curves that are missing in the zoom in view. Note that the zoom in view figures are also presented in Section 6 in the main text.



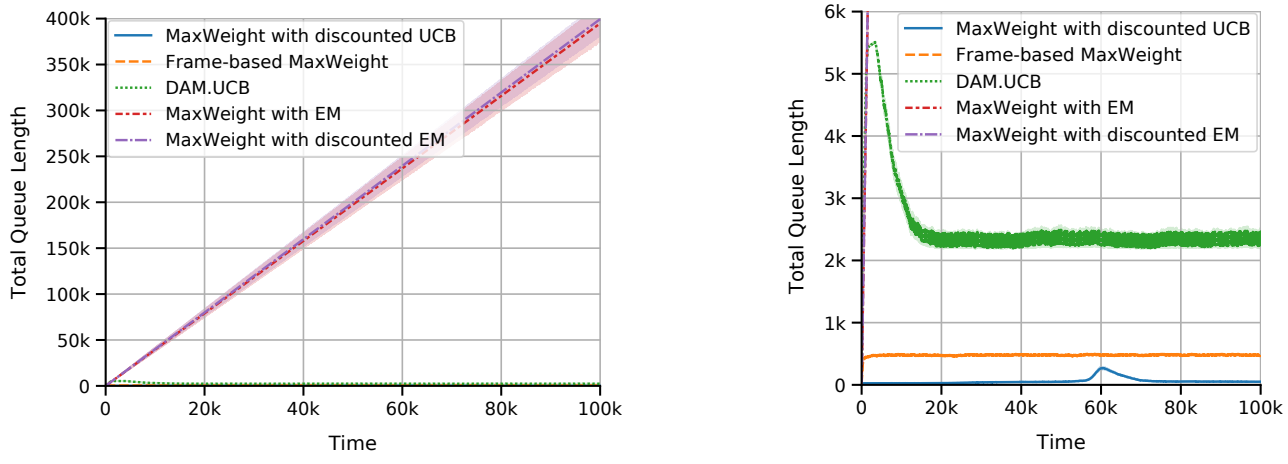
(a) Stationary.

(b) Nonstationary Aperiodic.

(c) Nonstationary Periodic.

(d) Nonstationary Periodic.

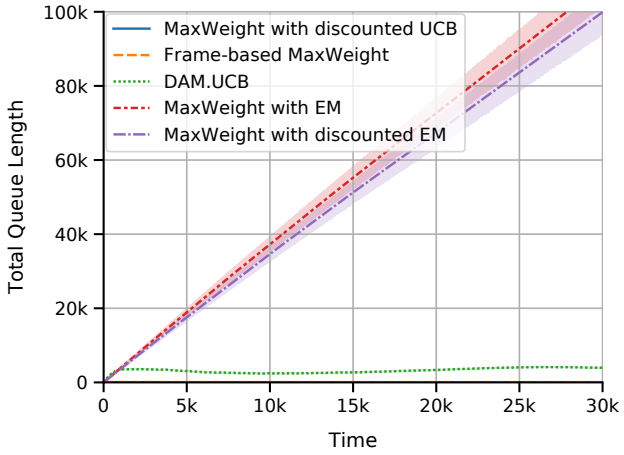
Figure 6: Simulation Results of *Frame-Based MaxWeight* With Different Frame Sizes.



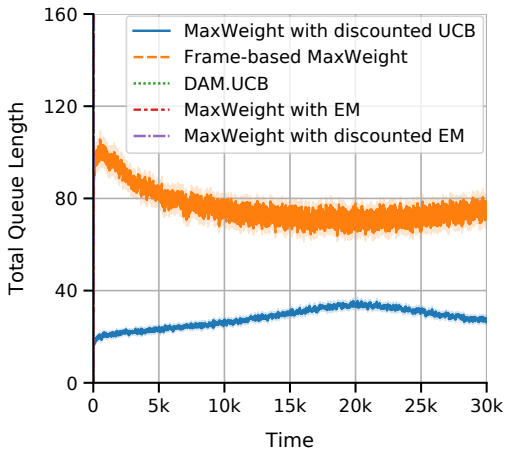
(a) Zoom Out View

(b) Zoom In View

Figure 7: Stationary Arrival Rate and Service Rate.

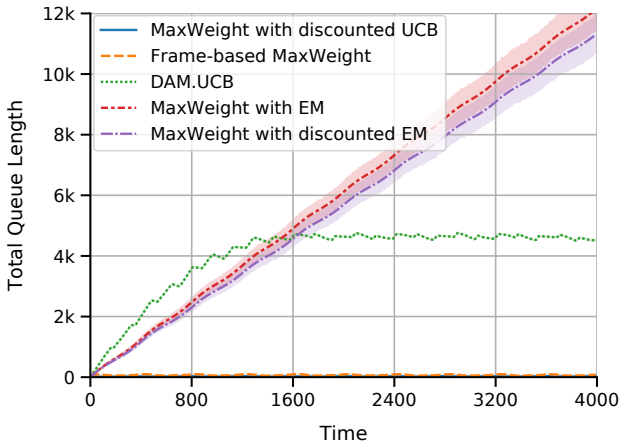


(a) Zoom Out View

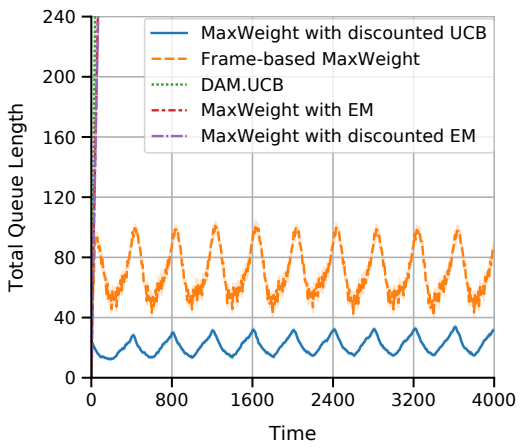


(b) Zoom In View

Figure 8: Nonstationary Service Rate With Aperiodic Means.

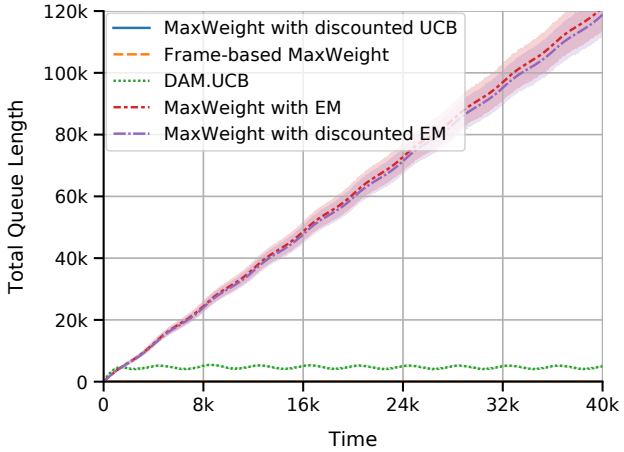


(a) Zoom Out View

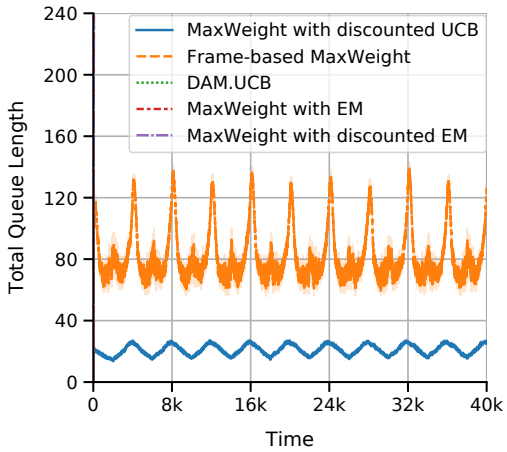


(b) Zoom In View

Figure 9: Nonstationary Arrival Rate and Service Rate With Periodic Means.



(a) Zoom Out View



(b) Zoom In View

Figure 10: Nonstationary Arrival Rate and Service Rate With Periodic Means With a Larger Period.