Online Job Assignment

Farbod Ekbatani
University of Chicago Booth School of Business, Chicago, fekbatan@chicagobooth.edu

Yiding Feng
University of Chicago Booth School of Business, Chicago, yiding.feng@chicagobooth.edu

Ian Kash
Computer Science, University of Illinois at Chicago, Chicago, iankash@uic.edu

Rad Niazadeh
University of Chicago Booth School of Business, Chicago, IL, rad.niazadeh@chicagobooth.edu

Abstract. Motivated primarily by applications in cloud computing marketplaces, we study a simple, yet powerful, adversarial online allocation problem in which jobs of varying durations arrive over continuous time and must be assigned immediately and irrevocably to one of the available offline servers. Each server has a fixed initial capacity, with assigned jobs occupying one unit for their duration and releasing it upon completion. The algorithm earns a reward for each assignment upon completion. We consider a general *heterogeneous* setting where both the reward and duration of a job assignment depend on the job-server pair. The objective of the online algorithm is to maximize the total collected reward while remaining competitive against an omniscient benchmark that knows all job arrivals in advance. Our main contribution is the design of a new family of online assignment algorithms, termed *Forward-Looking BALANCE (FLB)*, and using primal-dual framework to establish its competitive ratio. After proper selection of parameters, this family is (asymptotically) optimal-competitive in various regimes.

In summary, this meta-algorithm has two important primitives: (i) keeping track of the capacity used for each server at each time and applying a penalty function to this quantity, and (ii) adjusting the reward of assigning an arriving job to a server by subtracting the total penalty of a particularly chosen subset of future times (referred to as *inspection times*), in contrast to just looking at the current time. The FLB algorithm then assigns the arriving job to the server with the maximum adjusted reward. In the general setting, if R and D are the ratios of maximum over minimum rewards and durations, we show that there exists a choice of these primitives so that the FLB algorithm obtains a competitive ratio of $\ln(RD) + 3\ln\ln(R \vee D) + \mathcal{O}(1)$ as the initial capacities converge to infinity. Furthermore, in the special case of fixed rewards ($R = 1$), we show that the FLB algorithm with a different choice of primitives obtains a near-optimal asymptotic competitive ratio of $\ln(D) + \mathcal{O}(1)$. Our main analysis combines a dual-fitting technique, which leverages the configuration LP benchmark for this problem, and a novel inductive argument to establish the capacity feasibility of the algorithm, which might be of independent interest.

1. Introduction

The real-time allocation of available supply to incoming demand is crucial for efficient service operations and revenue management of many modern online platforms. Traditional applications, such as ad allocation (Karp et al., 1990; Mehta et al., 2007; Buchbinder et al., 2007; Aggarwal et al., 2011) and assortment planning (Golrezaei et al., 2014; Ma and Simchi-Levi, 2020), typically involve assignments where a demand request consumes a unit of supply for the entire decision-making horizon. For instance, in display advertising, once an impression is assigned to an ad, the advertiser's budget is immediately and irreversibly reduced by the bid amount. Similarly, in online retail, when a consumer purchases a product from the displayed assortment, that unit is permanently removed from the retailer's inventory. However, this paradigm shifts in modern e-commerce platforms and online marketplaces, such as cloud computing services like AWS and Microsoft Azure. In these environments, computing resources—such as virtual machines or memory—are allocated to jobs for their duration and are released upon completion, making them available for future tasks.

This novel feature of cloud platforms introduces emerging challenges in efficiently allocating computing resources in real time to users who compensate the platform for processing their jobs. Motivated primarily by this application, we study a simple, yet fundamental, sequential allocation problem, which we call *online job assignment*. In this problem, a platform aims to manage the assignment of jobs to available servers. Jobs arrive online in continuous time and must be immediately assigned to a compatible server for processing or rejected (no advance booking is allowed). Once assigned, a server must process the job continuously and without interruption (no preemption) for a certain "duration" of time. Each server has a capacity (of at least $c_{\min} \in \mathbb{N}$), which represents the maximum number of concurrent jobs that it can process at any given time. Upon assigning a job to a server, the platform receives a "reward" for the completion of each time unit of the job (which can represent monetary payments to the platform, throughput, or user satisfaction). The objective of the platform is to assign arriving jobs to servers in an online fashion, considering their per-unit rewards and durations, to maximize the total collected rewards while respecting the servers' capacity constraints.

We aim to design competitive online algorithms for our problem and compare them to the optimal offline benchmark, which has full knowledge of the entire set of arriving jobs (thus clearly providing an upper bound on the expected reward achievable by any online algorithm). Importantly, we consider the fully *heterogeneous* setting, where both the duration required to complete a job and the platform's reward may vary depending on the specific job and the server to which it is assigned. This is primarily motivated by the fact that different jobs may have distinct computational requirements, and servers typically differ in specifications, resulting in varied compatibilities regarding these requirements. Consequently, both the completion time of a job on a server and the reward generated from the assignment (e.g., the price paid by the user) may differ significantly.

Specifically, after normalizing the minimum (non-zero) reward and duration to one, we specify the heterogeneity of a given problem instance by the parameters $R \geq 1$ and $D \geq 1$, representing the maximum reward and maximum duration among all jobs, respectively. We say an algorithm is $\Gamma(R, D)$-*competitive* for some

$\Gamma(R,D)\geq 1$ if its total reward for any instance with heterogeneity parameters (R,D) is at least $\frac{1}{\Gamma(R,D)}$ times the total reward obtained by the optimal offline benchmark. Focusing on the practically relevant and widely studied "large capacity regime" or "asymptotic regime," where $c_{\min}$ is large (see, e.g., the classic AdWords problem in Mehta et al., 2007), we address the following research question regarding the competitive ratio and heterogeneity parameters R and D:

Question (i): *In the online job assignment problem, given heterogeneity parameters $R\geq 1$ (for rewards) and $D\geq 1$ (for durations), can we design an online algorithm—ideally simple and practical—with an asymptotically optimal competitive ratio $\Gamma(R,D)$?*

Based on the prior literature studying special cases of our model—in particular, (i) single or multiple non-reusable resources with heterogeneous rewards (Ball and Queyranne, 2009; Ma and Simchi-Levi, 2020; Feldman et al., 2009; Ekbatani et al., 2023), and (ii) single or multiple reusable resources with heterogeneous rewards and durations (Huo and Cheung, 2022; Rusmevichientong et al., 2023)—and from the literature on general online packing and covering models—in particular, (i) online routing, load balancing, and interval scheduling (Buchbinder and Naor, 2009; Awerbuch et al., 1993; Lipton and Tomkins, 1994; Azar et al., 1997), and (ii) offline combinatorial auctions (Buchbinder and Gonen, 2015)—we know that the asymptotic optimal competitive ratio should exhibit at least a logarithmic dependency on R and D, or more specifically, $\Gamma(R,D)=\Omega\left(\log(RD)\right)$.[1] See Sections EC.1 and 1.3 for details on these results for the mentioned special cases of our problem. Given these lower bounds, for any online algorithm $\mathcal{A}$ in the online job assignment problem, we define its *asymptotic logarithmic constant* $\mathcal{C}_{\text{LOG-LIMIT}}(\mathcal{A})$ as the smallest constant $C>0$ satisfying:

$$\Gamma(R,D)=C\cdot\ln(RD)+o\left(\log(RD)\right),$$

when the minimum server capacity $c_{\min}$ is sufficiently large. By convention, we set $\mathcal{C}_{\text{LOG-LIMIT}}(\mathcal{A})=+\infty$ if $\Gamma(R,D)=\omega\left(\log(RD)\right)$. We can now rephrase our research question as follows:

Question (ii): *In the online job assignment problem, can we characterize the optimal asymptotic logarithmic constant $\mathcal{C}^*_{\text{LOG-LIMIT}}$, defined as the smallest asymptotic logarithmic constant $\mathcal{C}_{\text{LOG-LIMIT}}(\mathcal{A})$ achievable by some online algorithm $\mathcal{A}$?*

1.1. Our Main Contributions

In this paper, we provide compelling answers to both of these questions. In a nutshell, we establish that $\mathcal{C}^*_{\text{LOG-LIMIT}}=1$, and that it can be achieved by a simple, interpretable, and practical algorithm. More formally, for the online job assignment problem, we show the following main result:

[1] Here, the standard *Bachmann–Landau notations* (or *asymptotic notations*)—including Big $\mathcal{O}$, small o, Big Ω, and small ω—are used to describe the limiting behaviors of our upper and lower performance ratio bounds.

[Main Result] *We propose a simple, practically relevant, and polynomial-time online algorithm (Algorithm 1) that attains the asymptotically optimal competitive ratio (up to lower-order terms) of* $\ln(R \cdot D)$ *for general choices of parameters* R *and* D *(and* $\frac{e}{e-1}$ *when* $R = D = 1$*), when the minimum server capacity* $c_{\min}$ *is sufficiently large. We further show that this bound is the best achievable by any online algorithm, thus establishing that* $\mathcal{C}^*_{\text{LOG-LIMIT}} = 1$*. See Table 1.*

Table 1 Summary of the main results (optimal asymptotic logarithmic constant of $\mathcal{C}^*_{\text{LOG-LIMIT}} = 1$).

		Upper Bound	**Lower Bound**
General ($R > 1, D > 1$)	Real Durations	$\ln(RD) + 3\ln\ln(R \vee D) + \mathcal{O}(1)$ **(Theorem. 3.1)**	$\ln(RD) + \Omega(1)$ **(Proposition. 5.1)**
	Integer Durations	$\ln(RD) + \ln\ln(R \vee D) + \mathcal{O}(1)$ **(Theorem. 4.1)**	(same as above)
Fixed-reward ($R = 1, D > 1$)	Real Durations	$\ln(D) + 4$ **(Proposition. EC.2.2)**	$\ln(D) + 2$ **(Proposition. EC.2.2)**
	Integer Durations	$H(D) + 2$ **(Proposition. 4.8)**	$H(D)$ **(Proposition. 4.8)**
Fixed-reward, Fixed-Duration ($R = 1, D = 1$)		$\frac{e}{e-1}$ **(Remark. 4.1)**	$\frac{e}{e-1}$ **Karp et al. (1990); Mehta et al. (2007)**

Note: $H(k) \triangleq \sum_{i=1}^{k} i^{-1} \in [\ln(k), \ln(k) + 1]$ *denotes the k-th harmonic number; operator* $\vee$ *is* max *and all* $\ln(\cdot)$ *are natural logarithms. The upper bound of* $\frac{e}{e-1}$ *is known in Feng et al. (2019); Goyal et al. (2025) for the setting of fixed reward, fixed duration and discrete-time arrival.*

Our paper extends and unifies the scope of several previous works in the literature on online resource allocation, each focusing on specific aspects of the problem. Within our comprehensive model, we achieve asymptotically near-optimal competitive ratios (with almost optimal dependencies on the parameters R and D). Our technical developments yield a significant improvement in terms of the asymptotic logarithmic constant over prior work concerning special cases of our problem, using a simple and practical algorithm. Moreover, we recover the exact asymptotic optimal bounds in the special cases corresponding to $R = 1$ and to $R = D = 1$, studied previously in Feng et al. (2019); Gong et al. (2022); Delong et al. (2024); Goyal et al. (2025). We also note that although our primary analysis addresses the base setting in which the heterogeneity parameters are independent of the server choice, all of our results naturally extend to the server-dependent heterogeneous setting where each server i has its own reward range R_i and duration range D_i, with $R \triangleq \max R_i$ and $D \triangleq \max D_i$. See Section EC.4.

Finally, in Section 6, we evaluate the numerical performance guarantee of our proposed algorithm both on the class of worst-case instances we identify in this paper, and on randomized instances beyond the worst-case. We compare the algorithm with other existing benchmarks in the literature in terms of their instance-wise competitive ratio. We observe that not only our algorithm both theoretically and numerically has the best performance in the worst-case instance, but also in almost all of our simulations, our algorithm outperforms the other benchmarks in most randomized instances.

1.2. Overview of the Techniques

Before elaborating on the algorithmic development that leads to our main result, it is helpful to provide some context from the literature. As a special case of our model, Kalyanasundaram and Pruhs (2000) introduce the simple and elegant BALANCE algorithm for the online bipartite b-matching problem involving non-reusable resources with 0-1 rewards. Subsequent works extend this idea to other variants of online resource allocation, such as those with non-reusable resources and essentially identical rewards over time, e.g., Mehta et al. (2007); Aggarwal et al. (2011) (corresponding to $R = 1$ and infinite-duration jobs), as well as to reusable resources with identical rewards and durations over time, e.g., Feng et al. (2019, 2021) (corresponding to $R = D = 1$).

The central idea behind the BALANCE algorithm is to evenly distribute the load over time by computing a reduced reward for each resource (in our setting, each server), based on the resource's current available capacity at each point in time. The algorithm then greedily assigns each arriving request (or job, in our setting) to the resource with the highest positive reduced reward. Notably, this reward adjustment can be interpreted through a primal-dual perspective, where the available capacity is used as a surrogate by the online algorithm to maintain an approximate dual solution, and then rewards are adjusted according to this dual solution following complementary slackness.

The BALANCE algorithm achieves the asymptotically optimal competitive ratio of $\frac{e}{e-1}$ when $R = D = 1$ in the large-capacity regime. However, it remains unclear whether this approach can be successfully applied or extended to the heterogeneous online job assignment setting (i.e., when $R > 1$ or $D > 1$). Indeed, we have both strong theoretical (see Example 3.3) and empirical (see Section 6) evidences suggesting that such an extension would fail to maintain the algorithm's performance guarantees, as it only considers the current server utilizations (equivalently, the available capacity), without accounting for how *future utilizations* may evolve. Our proposed algorithm aims to *precisely* address this limitation.

Forward-Looking BALANCE: "forecasting" future utilizations. Our algorithm is a novel generalization of BALANCE. We introduce a new parametric family of online algorithms, called Forward-Looking BALANCE (FLB), that extends the original idea of achieving a "balanced" allocation based on adjusted rewards to the online job assignment problem with arbitrary heterogeneity in rewards and durations of arriving jobs. To handle the reusability of servers and heterogeneity across jobs, a balanced allocation must consider not only the current available capacities of servers but also the (projected) future trajectories of these available capacities—thus being "forward-looking." Intuitively speaking, while the current number of jobs assigned to a server (reflected in its current available capacity) indicates how busy the server is, it provides an incomplete picture in settings with heterogeneous durations: currently assigned jobs will finish one by one at different future times, increasing the server's available capacity at each of those times. Accounting for this future capacity recovery can potentially improve current assignment decisions.

To illustrate this point, consider a simple scenario involving two servers with identical rewards and durations for a particular job, and imagine that the current available capacities of these servers are also identical.

In this situation, the classic BALANCE would calculate the same reduced reward for both servers and is thus indifferent to assigning the job to either of them. However, if server 1 is expected to regain all currently unavailable capacity units within the next second, while the unavailable units of server 2 become available much later or never (due to significantly longer-duration jobs assigned to this server in the past), it becomes intuitively preferable to prioritize server 1, which will soon have higher available capacity, over server 2. This forward-looking prioritization maintains balanced resource utilization over time. FLB formalizes this intuitive idea by incorporating both the current available capacity as well as the anticipated future available capacities into the calculation of the reduced reward for each server. For additional details, see the discussion in Section 3, particularly Example 3.3 and Figure 1.

Customizing the penalty function and inspection times. Given the inherent heterogeneity of the environment, particularly due to varying job durations, accurately capturing the trade-off between a server's immediate availability and its anticipated future availability through reward adjustments poses intricate challenges, necessitating novel algorithmic flexibility. As a first form of flexibility, we allow FLB to use the parameters R and D to determine how it adjusts rewards. Specifically, we employ the following *family of exponential penalty functions* $\Psi : [0,1] \longrightarrow \mathbb{R}_+$, parameterized by (β, η), to translate the (projected) normalized available capacity at each current and future inspection time into a dual-based penalty:

$$\Psi(x) = \eta \left(\beta^{(1-x)} - 1\right),$$

where parameters $\beta \geq 1$ and $\eta \geq 0$ may depend on R and D. As a second form of flexibility, FLB addresses duration heterogeneity by selecting only a (possibly infinite) subset of future times—referred to as *inspection times*—to anticipate the trajectory of available capacity when making the current assignment decision. At any given continuous time t, these inspection times for an arriving job on each server are periodic points at a fixed frequency $\gamma \geq 0$, chosen from the set of future times during which this job causes resource conflicts on that server (i.e., the interval $[t, t+d]$, where d is the job's duration on that server); see Section 3 for a formal definition. Similar to (β, η), the frequency parameter γ may depend on the heterogeneity parameters R and D. Ultimately, the reward adjustment at time t reduces the original reward by the sum of all dual-based penalties corresponding to current and future inspection times. For technical reasons elaborated further in Section 4, the choices of the penalty function and frequency of future inspections both play critical roles in our primal-dual analysis, which we briefly describe next.

Primal-dual analysis with configuration LP. We rely on a novel primal-dual analysis to establish the competitive ratios of FLB. Similar to prior works on variants and special cases of the online job assignment problem, we consider a linear programming (LP) relaxation $\mathcal{P}_{\text{OPT}}$ of the optimal offline policy and its corresponding dual program. However, unlike the standard LP relaxations used for primal-dual analysis in this literature, where each primal variable represents the probability of matching a job-server pair (Feng et al., 2019; Huo and Cheung, 2022), we introduce a *configuration linear program* (see $\mathcal{P}_{\text{OPT}}$) to help with a refined analysis of FLB and establish the (almost) optimal dependencies of the competitive ratio on both R and D.

In this configuration LP, each server is assigned to a collection of feasible configurations, where each feasible configuration is a subset of non-overlapping jobs that can all be processed using the same unit of capacity on that server. Each job is used in exactly one configuration, and each unit of server capacity is assigned precisely to one feasible configuration (the units are identical, so no server has a collection of feasible configurations larger than its capacity). Such a configuration linear program can encode any feasible assignment of jobs to servers given their capacities, regardless of duration heterogeneity. This granular, "higher-dimensional representation" of our assignment is one of the key technical insights that allows us to construct proper dual solutions, certifying the approximate optimality of the primal solution generated by FLB.

More specifically, in the dual program of our configuration LP (see $\mathcal{P}_{\text{OPT}}$), there is a non-negative dual variable associated with each job and server. The dual objective function is simply the sum of all these dual variables, while the dual constraints impose limitations on the sum of dual variables corresponding to each server and each subset of jobs that can simultaneously be assigned to the same unit of capacity on that server. The primal-dual method aims to construct a dual assignment satisfying two key properties: (i) its objective value is at most $\Gamma(R, D)$ times the total reward collected by FLB; and (ii) all dual constraints are satisfied. To achieve this, we explicitly construct a dual assignment guided by the execution of FLB. We establish property (i) using the specific closed-form expression the algorithm employs to compute the reduced reward and demonstrate property (ii) via a novel constructive charging argument.

As will become clear later in Section 4, proving property (ii)—and thus obtaining our refined bounds with near-optimal dependencies on R and D—is quite subtle. The key insight behind our improved analysis is to *not* force feasibility of the primal solution generated by FLB through an overly conservative choice of the penalty function (which sets the reduced reward to zero whenever the current capacity reaches zero). Instead, we adopt an alternative penalty function and show that feasibility naturally arises as an *indirect* consequence of running our algorithm. This reasoning relies on maintaining a specific invariant property throughout the algorithm's execution, representing another novel aspect of our approach, which we briefly discuss next.

Primal capacity feasibility using an invariant. In both the classic BALANCE and our proposed FLB, the penalty function must be carefully designed to ensure capacity feasibility. Analyzing the capacity feasibility of BALANCE for non-reusable resources is relatively straightforward, as it involves considering a single scenario where the entire capacity is exhausted. However, in our setting, verifying the capacity feasibility of FLB involves infinitely many scenarios—capacity can become temporarily exhausted, and future capacity levels vary across these scenarios. To guarantee the feasibility in all such scenarios, the algorithm designer might be forced to adopt a pessimistic penalty function, which, as previously discussed, would lead to poor competitive performance. To improve competitive performance, we introduce an *invariant property* that essentially characterizes the possible future capacity levels of FLB given different choices of the penalty function. This invariant allows us to exclude pessimistic scenarios where capacity feasibility could fail, as these scenarios will never occur in any execution path of FLB. To prove this invariant property, we develop a novel inductive argument that also clarifies the selection of specific parameters (γ, η, β) for FLB, and reveals how they are influenced by the parameters R and D.

1.3. Related Work

Our work is connected to various lines of literature in operations research and computer science. In addition to the related work discussed below, other related literature can be found in Section EC.1.

Online resource allocation of non-reusable resources. There is a long literature about online resource allocation, where the consumer sequence is determined by an adversary, known as the adversarial arrival setting. In the basic model of online bipartite matching, Karp et al. (1990) introduce the RANKING algorithm and demonstrate its optimality with a competitive ratio of $\frac{e}{e-1}$. Kalyanasundaram and Pruhs (2000) study the online bipartite b-matching where each offline node has a capacity (inventory) constraint. The authors propose the BALANCE algorithm and show it also achieves the optimal $\frac{e}{e-1}$-competitive under a large inventory assumption. Later work generalizes the BALANCE algorithm to various other models, including the AdWords problem (Mehta et al., 2007), the online assortment problem (Golrezaei et al., 2014), batch arrival (Feng and Niazadeh, 2025), and unknown capacity (Manshadi et al., 2025). All aforementioned works assume that each resource has an identical reward that is independent of the consumer. For online resource allocation of non-reusable resources with heterogeneous rewards, Ball and Queyranne (2009) investigate the same model as ours but with the restriction that all resources are non-reusable. They introduce a "protection level" policy and demonstrate its asymptotic optimality with a competitive ratio of $O(\log R)$ under a large inventory assumption. A similar result is also found in Buchbinder and Naor (2009) and subsequent works (e.g., Azar et al., 2016) that study the online packing/covering problem. Ma and Simchi-Levi (2020) study the same problem, present a generalization of the BALANCE algorithm, and show it achieves a more refined instance-dependent competitive ratio. Our results align with the competitive ratio results established in Kalyanasundaram and Pruhs (2000); Ball and Queyranne (2009); Ma and Simchi-Levi (2020).

Online resource allocation of reusable resources. Several studies have been conducted on online resource allocation of reusable resources in the adversarial arrival setting. Gong et al. (2022) study the adversarial setting with identical rental fees and stochastic i.i.d. rental duration and show that the greedy algorithm achieves a competitive ratio of 2. Simchi-Levi et al. (2025) introduce a more general model with decaying rental fees and rental duration distributions. Goyal et al. (2025) study the same model as Gong et al. (2022) and design a $\frac{e}{e-1}$ competitive algorithm under a large inventory assumption. Feng et al. (2019) and Delong et al. (2024) study the online bipartite matching of reusable resources with identical rental fees and identical rental durations. Under a large inventory assumption, Feng et al. (2019) show that the BALANCE algorithm is $\frac{e}{e-1}$ competitive. Delong et al. (2024) design a 1.98-competitive algorithm without a large inventory assumption.

All the aforementioned previous works consider identical rental fees and durations, without considering heterogeneity. In contrast, our paper addresses the issue of heterogeneous rental fees and durations. Notably, all algorithms in these prior works are not forward-looking, i.e., the allocation decision is made based on the current inventory without an eye toward the anticipated inventory level in the future. Our result recovers the result in Feng et al. (2019). Finally, we note the growing body of work that considers the learning problem in the allocation and pricing of reusable resources Jia et al. (2024); Feng et al. (2024). This differs from our work, as we focus on the adversarial setting rather than a stochastic learning framework.

A recent closely related work is Huo and Cheung (2022). This paper shares the same model as ours but introduces the additional restriction of having only a single resource. They further make three types of monotonicity assumptions regarding consumers' personalized rental fees and durations. They generalize the protection level policy proposed by Ball and Queyranne (2009). Under a large inventory assumption, they prove a competitive ratio of $\phi \cdot \ln(RD)$, $\phi \cdot R\ln(RD)$, and $\phi \cdot D\ln(RD)$ for three types of monotonicity assumptions respectively, where $\phi \in [1, 2)$ is a constant that depends on the consumer sequence. In comparison to Huo and Cheung (2022), our paper considers multiple reusable resources without imposing any monotonicity assumptions and achieves an asymptotically optimal competitive ratio of $\ln(RD)$. Therefore, the competitive ratio of $\ln(RD)$ in our paper not only improves the coefficient in a more general model but also represents an exponential improvement compared to Huo and Cheung (2022). Finally, it should be noted that Huo and Cheung (2022) uses the standard LP as the benchmark for the analysis of competitive ratios. Consequently, their primal-dual proof requires the aforementioned monotonicity assumption, and the dual assignment construction necessitates the utilization of a novel, albeit complex, auxiliary algorithm. In contrast, our paper considers the competitive ratio with respect to the configuration LP, enabling a relatively simpler primal-dual analysis and achieving an asymptotically optimal competitive ratio without any monotonicity assumptions.

Online covering/packing LPs under adversarial arrivals. Another closely related line of work is the classic literature in computer science on online covering and online packing linear programs under adversarial arrivals (Buchbinder and Naor, 2009), and the related models such as online routing (Awerbuch et al., 1993), online load balancing (Azar et al., 1997), interval scheduling (Lipton and Tomkins, 1994), and combinatorial auction (Buchbinder and Gonen, 2015). By using non-trivial arguments, one might be able to show that special cases of the online job assignment problem with discrete time, integer durations, and single server can be reduced to these models, resulting in (somewhat complicated) algorithms that obtain $\mathcal{O}(\log(R \cdot D))$ competitive ratios in these special cases (with highly sub-optimal asymptotic logarithmic constants). This is completely in contrast to our simple and practical FLB algorithm, which obtains the optimal asymptotic logarithmic constant of $\mathcal{C}^*_{\text{LOG-LIMIT}} = 1$ in the most *general* case of the online job assignment problem with continuous time, real durations and multiple servers.

2. Problem Formulation

We study the *online job assignment* problem in a cloud computing platform, where jobs arrive online and servers are available offline. In this setting, there are n servers, indexed by $[n] \triangleq \{1,2,\dots,n\}$. Each server i has a capacity denoted by $c_i \in \mathbb{N}$, representing the maximum number of jobs the server can simultaneously process at any given time. We consider a finite continuous-time horizon $[0,T]$, during which m jobs arrive at times $0 \le t_1 \le \cdots \le t_m \le T$. The set $E \subseteq [n] \times [m]$ encodes the compatibility between servers and jobs. Each edge $(i,j) \in E$ is associated with a per-period reward $r_{ij} \in \mathbb{R}_+$, representing the reward the platform receives per unit time if job j is assigned to server i. Additionally, each edge $(i,j) \in E$ has an associated duration $d_{ij} \in \mathbb{R}_+$, representing the time required for job j to complete on server i (which may differ across servers).

Upon the arrival of job j, the set of compatible servers $N(j) \triangleq \{i \in [n] : (i,j) \in E\}$ and their corresponding rewards and durations $\{r_{ij}, d_{ij}\}_{i \in N(j)}$ are revealed to an online algorithm. The algorithm then immediately and irrevocably assigns job j to one of the compatible servers with available capacity, if any. We refer to the tuple $(N(j), \{r_{ij}, d_{ij}\}_{i \in N(j)})$ as the type of job j. The platform receives a total reward of $r_{ij} d_{ij}$ for assigning (and completing) job j on server i. The completion time $t_j + d_{ij}$ may exceed the time horizon T, in which case the server's capacity remains occupied until time T. The goal of the online algorithm is to maximize the total reward collected during the time interval $[0,T]$. We highlight that this model is similar to the online matching with reusable resources (Udwani, 2025; Feng et al., 2019), but we diverge in that the rewards and durations are *heterogeneous over time* (and arrivals occur in continuous time).

We make the following regularity assumption to bound the heterogeneity of rewards and durations, without which meaningful performance guarantees are not possible (Azar et al., 2015).

Assumption 2.1 (Bounded Heterogeneity) *For every job $j \in [m]$ and every compatible server $i \in N(j)$, the associated reward r_{ij} and duration d_{ij} satisfy $r_{ij} \in [1,R]$ and $d_{ij} \in [1,D]$, where R and D denote the maximum reward and maximum duration across all job-server pairs, respectively.*

Remark 2.2 (Known Parameters R and D) *We assume that the online algorithm knows the ranges of rewards R and durations D. This assumption is well-motivated in practical applications and is also theoretically necessary. Specifically, Section EC.5, together with Theorem 3.1, shows a strict separation in performance between algorithms with and without knowledge of (R,D). Similar assumptions have also been previously made in the literature for special cases of our problem (e.g., Buchbinder and Naor, 2009; Chawla et al., 2019; Ma and Simchi-Levi, 2020).*

Given parameters $R \ge 1$ and $D \ge 1$, we evaluate the performance of an online algorithm by its *competitive ratio* against the optimal offline benchmark with full knowledge of the sequence of job types in advance. Specifically, the competitive ratio is the worst-case ratio of the total reward obtained by the online algorithm to that of the optimal offline benchmark within instances with rewards at most R and durations at most D.

Throughout this work, we focus mainly on the *large capacity regime* when analyzing competitive ratios, where $c_{\min} \triangleq \min_{i\in[n]} c_i$ tends to infinity.[2]

More formally, letting $\mathcal{I}(R, D, c_{\min})$ denote the subset of problem instances with $1 \le r_{ij} \le R$ and $1 \le d_{ij} \le D$ for $j \in [m], i \in N(j)$,[3] and minimum capacity $c_{\min}$, we have the following definition.

Definition 2.3 (Asymptotic Competitive Ratio) *Given any $R, D \ge 1$, a (possibly randomized) online algorithm ALG is said to be "(asymptotically) $\Gamma(R, D)$-competitive" if the following holds:*

$$\limsup_{c_{\min}\to\infty} \left(\sup_{I\in\mathcal{I}(R,D)} \frac{\mathrm{OPT}(I)}{\mathbb{E}\left[\mathrm{ALG}(I)\right]} \right) \le \Gamma(R, D)$$

where $\mathrm{ALG}(I)$ and $\mathrm{OPT}(I)$, respectively, are the total rewards of the online algorithm ALG and the optimal offline benchmark OPT under the problem instance I.

3. Forward-Looking BALANCE

Our main result is a new online algorithm, known as Forward-Looking BALANCE (FLB). In this section, we describe this algorithm and all its ingredients. The competitive ratio analysis is deferred to Sections EC.2 and 4 (upper bound) and Section 5 (lower bound).

The main new idea behind FLB is simple and natural: When a job j arrives, the assignment decision for this job is made not only based on the type of job j and the current available capacities of compatible servers, but also based on the *projected available capacities* of each compatible server i at the current time t_j for each future time $\tau \ge t_j$ (or a subset of these times).

Definition 3.1 (Projected Available Capacity) *For each server $i \in [n]$ and time points $t, \tau \in [0, T]$ with $t \le \tau$, the (normalized) projected available capacity $\alpha_{i,t\to\tau} \in [0, 1]$ is defined as*

$$\alpha_{i,t\to\tau} \triangleq 1 - \frac{1}{c_i} \sum\nolimits_{j\in[m]:t_j<t} \mathbb{1}\left\{\mathcal{E}(i, j)\right\} \cdot \mathbb{1}\left\{t_j + d_{ij} > \tau\right\},$$

where $\mathcal{E}(i, j)$ is the event that a capacity unit of server i is assigned to job j.

In other words, the projected available capacity $\alpha_{i,t\to\tau}$ represents the proportion of available units for server i at time τ, taking into account the job assignments made before time t. For any assignment decisions, the resulting $\alpha_{i,t\to\tau}$ is weakly increasing in τ and weakly decreasing in t. Furthermore, $\alpha_{i,t\to t}$ denotes the proportion of available units for server i at time t.

[2] This regime is practically relevant in cloud marketplaces and standard in theoretical analyses within the online matching literature, e.g., Kalyanasundaram and Pruhs (2000); Mehta et al. (2007); Golrezaei et al. (2014); Feng et al. (2019); Ma and Simchi-Levi (2020).

[3] In Assumption 2.1, we normalize both the minimum reward and the duration to 1 for compatible server-job pairs. This is without loss of generality (particularly for durations, since we consider a continuous-time horizon), and our results naturally extend to a more general setting, where each server $i \in [n]$ is associated with parameters $(\underline{r}^{(i)}, \underline{d}^{(i)})$, and the reward (resp. duration) range between server i and its compatible jobs $j \in N^{-1}(i)$ is $[\underline{r}^{(i)}, R \cdot \underline{r}^{(i)}]$ (resp. $[\underline{d}^{(i)}, D \cdot \underline{d}^{(i)}]$). See Section EC.4 for a detailed discussion.

To use the projected available capacities in its assignments, FLB is equipped with two important technical ingredients: *(i) inspection-time subset* and *(ii) projected-utilization-based reduced reward*.

(i) Inspection-time subset. As its name suggests, FLB makes job assignment decisions by taking a forward-looking perspective and examining the available capacity in the future. Specifically, at the arrival time t_j of each job j, FLB computes the projected available capacity $\alpha_{i,t_j\to\tau}$ for every inspection time $\tau\in\mathcal{T}_{ij}(\gamma)$ and every server $i\in[n]$. Here, the inspection-time subset $\mathcal{T}_{ij}(\gamma)\subseteq[t_j,t_j+d_{ij})$, parameterized by an *inspection-frequency scalar* $\gamma\in[1,\infty)$, is a subset of current and future time points defined as

$$\mathcal{T}_{ij}(\gamma)\triangleq\left\{\tau\in[t_j,t_j+d_{ij}):\exists\ell\in\mathbb{N}\text{ s.t. }\tau=t_j+\tfrac{\ell}{\gamma}\right\}.$$

Recall that when a unit of server i is assigned to job j, it remains occupied from time t_j and becomes available again at the beginning of time t_j+d_{ij}. Therefore, it is natural for the algorithm to examine the projected available capacity $\alpha_{i,t_j\to\tau}$ at inspection times $\tau\in\mathcal{T}_{ij}(\gamma)\subseteq[t_j,t_j+d_{ij})$.

(ii) Projected-utilization-based reduced reward. Due to future uncertainty and job heterogeneity, it may be preferable to assign jobs less frequently to servers with high utilization (or equivalently, low available capacity) at the current time, or those projected to have high utilization at a future time, unless the immediate reward is substantial. To capture this inherent trade-off, we introduce the (projected-utilization-based) reduced reward for each job j and server $i\in N(j)$, defined as:

$$r_{ij}d_{ij}-\sum\nolimits_{\tau\in\mathcal{T}_{ij}(\gamma)}\Psi\left(\alpha_{i,t_j\to\tau}\right),$$

where $\Psi:[0,1]\to\mathbb{R}_+$ is a weakly decreasing and convex *penalty function* satisfying $\Psi(1)=0$. As a sanity check, when all units of server i are available for processing job j throughout the entire time horizon, the reduced reward equals the original reward. Furthermore, the reduced reward decreases as fewer units become available. Notably, we make *no* explicit assumptions on $\Psi(0)$, and therefore we do not enforce the reduced reward of a server with no available units to be negative. We also focus on a specific form of exponential penalty function Ψ with a varying base, parameterized by *penalty parameters* (η,β), where $\beta\geq1$ and $\eta\geq0$:

$$\Psi(x)=\eta\left(\beta^{(1-x)}-1\right).$$

Similar to the classic BALANCE algorithm (Kalyanasundaram and Pruhs, 2000; Mehta et al., 2007), FLB computes a score for all compatible servers upon the arrival of each job j. It then makes a greedy-style decision, assigning job j to the server i_j^* with the highest positive score, if such a server exists (otherwise, the job remains unassigned). The key difference is that FLB uses the projected-utilization-based reduced rewards as its scores, while the scores in BALANCE are based solely on current utilization. See Algorithm 1 for a formal description of the algorithm.

Importantly, we do not enforce feasibility constraints (due to servers having limited capacity) *explicitly* within FLB by controlling $\Psi(0)$. Instead, as highlighted in the next remark, its feasibility is achieved *implicitly*

Algorithm 1: Forward-Looking BALANCE (FLB)

Input: inspection-frequency scalar γ, penalty parameters (η, β)

1 **for** *each job* $j = 1$ *to* m **do**
  /* job j with type $(N(j), \{r_{ij}, d_{ij}\}_{i \in N(j)})$ arrives */
2   **if** $\max_{i \in N(j)} r_{ij} d_{ij} - \sum_{\tau \in \mathcal{T}_{ij}(\gamma)} \Psi\left(\alpha_{i,t_j \to \tau}\right) > 0$ **then**
3   let $i_j^* \leftarrow \arg\max_{i \in N(j)} \; r_{ij} d_{ij} - \sum_{\tau \in \mathcal{T}_{ij}(\gamma)} \Psi\left(\alpha_{i,t_j \to \tau}\right)$
4   assign job j to server i_j^*

through the structure of the algorithm and the appropriate parameter choices. This approach is one of the key insights underlying our design, enabling FLB to be a polynomial-time algorithm with correct dependence of the competitive ratio on R and D.

Remark 3.2 (Capacity Feasibility and Running Time) *In later sections, we identify conditions (Propositions 4.2 and 4.10) on parameters (γ, η, β) that ensure the "capacity feasibility" of* FLB*, meaning it never assigns a job to a server without available capacity by following lines (2-4) in Algorithm 1. Moreover,* FLB *can be implemented in polynomial time, even when its inspection-time subset $\mathcal{T}_{ij}(\gamma)$ is large or continuous, because there are only a polynomial number of time points at which the value of the projected available capacity changes.*

Comparison with BALANCE. If we set the inspection-frequency scalar $\gamma = 1$ for instances with maximum duration $D = 1$, the inspection-time subset $\mathcal{T}_{ij}(1)$ reduces to a singleton $\{t_j\}$. This implies that the reduced reward is computed solely based on the available capacity $\alpha_{i,t_j \to t_j}$ at the current time t_j, causing FLB to coincide with BALANCE. However, for general instances with $D > 1$, FLB may behave differently from BALANCE, as it also takes projected utilization levels at future times into account. The following example illustrates that this distinction plays a key role in enabling FLB to potentially outperform BALANCE.

Example 3.3 *Consider an example with $n = 2$ servers $\{1, 2\}$, each having capacity $c_i = 4$. At time $t = 4$, new job j arrives with durations $d_{1j} = d_{2j} = 2$ and rewards $r_{1j} = 1$, $r_{2j} = 1 + \epsilon$. The state of the servers before assigning job j is illustrated in Figure 1: server 1 (on the left) is processing two jobs ending at $\{4.8, 5\}$, while server 2 (on the right) is handling two jobs completing at $\{5.2, 6.3\}$. Since both servers have identical current available capacity,* BALANCE *prioritizes the higher immediate reward and assigns job j to server 2. However,* FLB *penalizes server 2 for its lower projected available capacity in the near future, and therefore assigns job j to server 1 instead. This choice better "hedges" against the risk of future (adversarial) job arrivals by maintaining more balanced future utilization. For instance, in the scenario where three jobs represented by the dashed rectangles in Figure 1 arrive in the future, both servers would retain sufficient available capacity to accept all three under* FLB*. This is indeed needed if these jobs are compatible with only one of the servers and generate high rewards.*

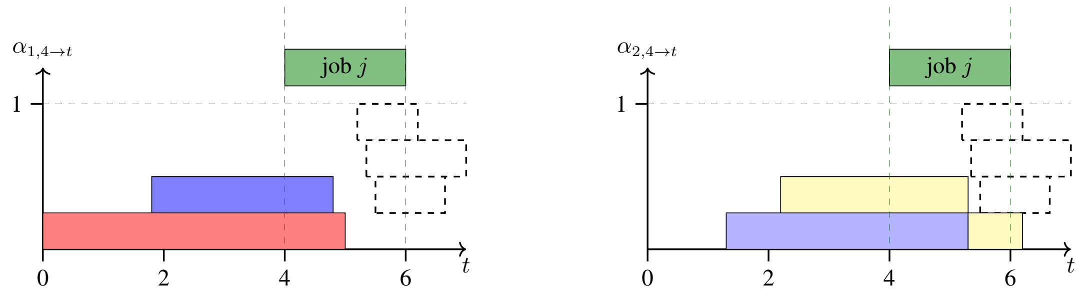

Figure 1 Graphical illustration of Example 3.3: Projected available capacity of the two servers at time $t = 4$ before new job j is assigned.

While the previous example and figure illustrated the potential benefits of incorporating projected future utilizations into an online algorithm, we now turn to the main theoretical result of this paper. Specifically, we demonstrate that, with an appropriate choice of parameters, FLB achieves an asymptotically optimal competitive ratio.

Theorem 3.1 (Competitive Ratio) *There exists a choice of parameters* $(\gamma^*, \eta^*, \beta^*)$ *such that the asymptotic competitive ratio of* FLB*, with inspection-frequency scalar* γ^* *and penalty parameters* (η^*, β^*)*, is at most*

$$\ln(RD) + 3\ln\ln(R \vee D) + O(1).$$

Moreover, there exist instances involving a single server with $R, D \geq 1$*, for which the asymptotic competitive ratio of any online algorithm (possibly fractional or randomized) against the optimal offline benchmark is at least* $\ln(RD) + \Omega(1)$. *As a corollary,* $\mathcal{C}_{\text{LOG-LIMIT}}(\mathsf{FLB}) = \mathcal{C}^*_{\text{LOG-LIMIT}} = 1$.

We remark that there is an explicit construction for $(\gamma^*, \eta^*, \beta^*)$ in the above theorem (and a closed form for these parameters when $c_{\min}$ approaches infinity in Section EC.2.3), which we detail later.

The proof of the upper bound result in the above theorem is deferred to Section EC.2, and the lower bound result is proved in Section 5. We also note that, although the theorem above is stated under the large-capacity assumption, our analysis additionally characterizes the non-asymptotic competitive ratio for any given capacity (see Propositions 4.3 and 4.11).

In the next section, we focus on proving a similar upper bound on the competitive ratio in a practical special case of our problem with integer-valued durations, which is slightly stronger than the upper bound in Theorem 3.1. Although this analysis relies on similar techniques, it simplifies the exposition of the main ideas used in the proof of Theorem 3.1. Later in Sections EC.2 and 4.4, we explain additional ideas needed in the upper bound analysis of FLB in the general case with real-valued durations.

4. Competitive Ratio Analysis of Integer-valued Durations

Consider a special case of our problem where, for every job $j \in [m]$ and compatible server $i \in N(j)$, the associated duration takes an integer value $d_{ij} \in \mathbb{N}$. We refer to this setting as the *integer-duration environment*.

This setting is practically motivated by scenarios in which the processing time for each job is a multiple of a minimum allowed processing interval,[4] and hence, after normalizing the minimum duration to one (without loss of generality), we can assume durations to be integers.[5]

Besides its practical relevance, this special case simplifies exposition of the core technical ideas required for the general case. Specifically, in this integer-duration environment, FLB can simply choose the inspection-frequency scalar $\gamma = 1$ and still achieve the asymptotically optimal competitive ratio, which simplifies the analysis. This contrasts with the general setting of real-valued durations, in which the algorithm must carefully select the inspection-frequency scalar γ.

Theorem 4.1 (Competitive Ratio for Integer-valued Durations) *In the integer-duration environment, there exists a choice of parameters (η^*, β^*) such that the asymptotic competitive ratio of* FLB*, with inspection-frequency scalar $\gamma^* = 1$ and penalty parameters (η^*, β^*), is at most*

$$\ln(RD) + \ln\ln(R \vee D) + O(1)$$

Similar to Theorem 3.1, the improved competitive ratio presented in Theorem 4.1 is asymptotically optimal, as the competitive ratio lower bound established in Theorem 3.1 remains valid for the integer-duration environment. An explicit construction (along with asymptotic closed-form parameter choices) for (η^*, β^*) to achieve the upper bound stated above is detailed in Section 4.3.

Overview of the analysis. In the remainder of this section, we present the formal proof of Theorem 4.1. At a high level, our analysis follows a three-step approach. Recall that FLB is parameterized by penalty parameters (β, η) and the inspection-frequency scalar γ. In the first step (Section 4.1), we characterize the set of parameter choices that guarantee the capacity feasibility of FLB, i.e., ensuring that the algorithm never assigns a job to a server lacking available capacity.

Proposition 4.2 (Capacity Feasibility of FLB for Integer-valued Durations) *In the integer-duration environment,* FLB *with inspection-frequency scalar $\gamma = 1$ is capacity feasible if penalty parameters (η, β) satisfy:*[6]

$$\ln(\beta) \geq -\ln\left(\prod\nolimits_{k\in[D]}\left(1 - \frac{R}{k(R+\eta)}\right) - \frac{(R+\eta)\ln(\beta)}{Rc_{\min}}\right)$$

In the second step (Section 4.2), we introduce a configuration LP, which upper bounds the optimum offline benchmark. We then use a primal-dual analysis based on this configuration LP to express the competitive ratio of the capacity feasible FLB as a function of its parameters.

[4] This scenario is common in cloud marketplaces, where usage is typically sold or charged in fixed time increments.

[5] While durations d_{ij} are assumed to be integer-valued, we still allow each job's arrival time t_j and reward r_{ij} to be real-valued.

[6] In fact, as we show in the proof of Proposition 4.2, this sufficient condition for the capacity feasibility becomes necessary as well when $c_{\min}$ approaches infinity.

Proposition 4.3 (Competitive Ratio of Capacity Feasible FLB for Integer-valued Durations) *In the integer-duration environment, for every $\eta > 0$ and $\beta \geq e$, the competitive ratio of a capacity feasible* FLB *with inspection-frequency scalar $\gamma = 1$ is at most*

$$\ln(\beta) \cdot \left(1 + \eta \left(1 + \beta \left(\beta^{\frac{1}{c_{\min}}} - 1\right)\right)\right)$$

Combining the two previous steps, the third step of our analysis (Section 4.3) formulates the problem of determining the best (asymptotic) competitive ratio of FLB as a constrained optimization problem over parameters (η, β), while fixing $\gamma = 1$. In Section 4.3, we formally present this optimization problem and analytically solve it in the large capacity regime (i.e., $c_{\min} \to \infty$). The solution for finite minimum capacity, $c_{\min} < \infty$, is provided in Section EC.6. These analytical solutions, along with a simple continuity argument, complete the proof of Theorem 4.1.

We highlight that the proof of Theorem 3.1 for general instances (with real-valued durations) follows the same structure as the proof of Theorem 4.1. However, both the derivation of sufficient conditions for capacity feasibility and the primal-dual analysis used to establish the competitive ratio become more involved; see Section EC.2 for details.

4.1. Sufficient Condition for the Capacity Feasibility

The ultimate goal of this section is to prove Proposition 4.2. To this end, we show that the current (normalized) available capacity $\alpha_{i,t\to t}$ remains non-negative throughout the execution of FLB if the parameters (η, β) satisfy the condition in the statement of Proposition 4.2.

As we illustrated in Example 3.3, a good online algorithm in our problem (such as FLB) should aim to evenly distribute resource utilization over time; in this way, it can avoid overcommitting to short jobs when the current utilization is high, thereby preserving the capacity for longer jobs that generate higher rewards. An algorithm can materialize this idea by maintaining a smooth trajectory for the projected available capacity $\alpha_{i,t\to\tau}$ as a function of τ. Specifically, we would like our algorithm to have control over the projected available capacity's variational terms $\alpha_{i,t\to\tau+d} - \alpha_{i,t\to\tau}$ for any $d \in \mathbb{N}$. We will formalize this property in the following lemma—which is maintained by FLB as an *invariant* throughout its execution. This invariant property is the key to characterize the set of parameters (β, η) that result in the capacity feasibility of FLB.

Lemma 4.4 (Invariant for Projected Available Capacities) *Consider a hypothetical scenario in which* FLB *can assign a job to a server that has no available capacity (which could result in a negative projected capacity level). In the integer-duration environment with inspection-frequency scalar $\gamma = 1$, the projected available capacity under* FLB *satisfies the following property: for every server $i \in [n]$, time points $t, \tau \in \mathbb{R}_+$ such that $\tau - t \in \mathbb{N}_+$, and integer duration $d \in \mathbb{N}$,* [7]

$$(\alpha_{i,t\to\tau+d} - \alpha_{i,t\to\tau}) \cdot \ln(\beta) \leq -\ln\left(\prod_{k\in[d]}\left(1 - \frac{R}{k(R+\eta)}\right) - \frac{(R+\eta)\ln(\beta)}{Rc_{\min}}\right).$$

[7] Throughout this section we choose β, η such that $\prod_{k\in[d]}\left(1 - \frac{R}{k(R+\eta)}\right) \geq \frac{(R+\eta)\ln(\beta)}{Rc_{\min}}$, which is possible for large enough $c_{\min} > 0$.

Proof. We prove the lemma by an induction over the sequence of currently arrived jobs $[0:j]$.

Base case ($j=0$): Before the first job arrives, the projected available capacity $\alpha_{i,0\to\tau}$ is 1 for all servers $i\in[n]$ and all future time points $\tau\geq 0$ by definition. Therefore, the inequality in lemma statement holds trivially as the left-hand side is zero and the right-hand side is non-negative (for large enough $c_{\min}$).

Inductive step ($j\geq 1$): Fix an arbitrary server $i\in[n]$ and suppose the inequality in the lemma statement holds for the arrival of first $j-1$ jobs. Now consider the arrival of job j at time t_j with duration d_{ij} for server i. We show the inequality holds after FLB determines the assignment for job j. The only non-trivial case happens when $t=t_j\leq\tau<t+d_{ij}\leq t+d$ and FLB assigns job j to server i. In this case, the reduced reward of server i is strictly positive, that is, $r_{ij}d_{ij}-\sum_{\tau'\in\mathcal{T}_{ij}(1)}\Psi(\alpha_{i,t_j\to\tau'})>0$. Combining the facts that $R\geq r_{ij}$ and $\mathcal{T}_{ij}(1)=\{t_j+k:k\in[0:d_{ij}-1]\}$ in the integer-duration environment, we obtain

$$\begin{aligned}
\frac{1}{\eta}Rd_{ij} &> \frac{1}{\eta}\sum\nolimits_{k'\in[0:d_{ij}-1]}\Psi\left(\alpha_{i,t_j\to t_j+k'}\right)\\
&=\frac{1}{\eta}\sum\nolimits_{k'\in[0:\tau-t_j-1]}\Psi\left(\alpha_{i,t_j\to t_j+k'}\right)+\frac{1}{\eta}\sum\nolimits_{k'\in[0:d_{ij}+t_j-\tau-1]}\Psi\left(\alpha_{i,t_j\to\tau+k'}\right)\\
&\geq(\tau-t_j)\left(\beta^{(1-\alpha_{i,t_j\to\tau})}-1\right)+\sum\nolimits_{k'\in[0:d_{ij}+t_j-\tau-1]}\left(\beta^{\left(1-\alpha_{i,t_j\to\tau+k'}\right)}-1\right)\\
&\overset{(a)}{\geq}-d_{ij}+(\tau-t_j)\,\beta^{\left(1-\alpha_{i,t_j\to\tau}\right)}+\sum\nolimits_{k'\in[0:d_{ij}+t_j-\tau-1]}\beta^{1-\alpha_{i,t_j\to\tau}+\frac{\ln\left(\prod_{k\in[k']}\left(1-\frac{R}{k(R+\eta)}\right)-\frac{(R+\eta)\ln(\beta)}{Rc_{\min}}\right)}{\ln(\beta)}}\\
&\geq-d_{ij}+\beta^{(1-\alpha_{i,t_j\to\tau})}\sum\nolimits_{k'\in[0:d_{ij}-1]}\left(\prod\nolimits_{k\in[k']}\left(1-\frac{R}{k(R+\eta)}\right)-\frac{(R+\eta)\ln(\beta)}{Rc_{\min}}\right)\\
&\overset{(b)}{=}-d_{ij}+\beta^{(1-\alpha_{i,t_j\to\tau})}\left(\frac{d_{ij}}{\frac{\eta}{R+\eta}}\prod\nolimits_{k\in[d_{ij}]}\left(1-\frac{R}{k(R+\eta)}\right)-\frac{d_{ij}(R+\eta)\ln(\beta)}{Rc_{\min}}\right)
\end{aligned}$$

where inequality (a) holds due to the induction hypothesis and the fact that $\tau-t\leq d_{ij}$, and equality (b) holds due to the following identity (with proof provided in Section EC.3.1).

Lemma 4.5 (Identity) *For any $d\in\mathbb{N}$ and $z\geq 0$, $\sum_{\ell\in[0:d-1]}\prod_{k\in[\ell]}(1-\frac{z}{k})=\frac{d}{1-z}\prod_{k\in[d]}(1-\frac{z}{k})$.*

Now, after moving d_{ij} to the left hand side and canceling $\left(\frac{R}{\eta}+1\right)d_{ij}$ from both sides, and then taking a logarithm from both sides and rearranging the terms, we have:

$$-\ln\left(\prod\nolimits_{k\in[d_{ij}]}\left(1-\frac{R}{k(R+\eta)}\right)-\frac{\eta\ln(\beta)}{Rc_{\min}}\right)\geq\left(1-\alpha_{i,t_j\to\tau}\right)\cdot\ln(\beta)\ .$$

and consequently,

$$\begin{aligned}
&-\ln\left(\prod\nolimits_{k\in[d]}\left(1-\frac{R}{k(R+\eta)}\right)-\frac{(R+\eta)\ln(\beta)}{Rc_{\min}}\right)\\
&\overset{(a)}{\geq}-\ln\left(\prod\nolimits_{k\in[d_{ij}]}\left(1-\frac{R}{k(R+\eta)}\right)-\frac{\eta\ln(\beta)}{Rc_{\min}}\right)+\frac{\ln(\beta)}{c_{\min}}\\
&\overset{(b)}{\geq}\left(1-\left(\alpha_{i,t_j\to\tau}-\frac{1}{c_{\min}}\right)\right)\cdot\ln(\beta)
\end{aligned}$$

$$\overset{(c)}{\geq} \left(\alpha_{i,t\to\tau+d} - \left(\alpha_{i,t\to\tau} - \frac{1}{c_i}\right)\right) \cdot \ln(\beta) \ .$$

where inequality (a) holds due to the concavity of the logarithm function and the case assumption that $d \geq d_{ij}$, inequality (b) holds as we argued above, and inequality (c) holds since $\alpha_{i,t\to\tau+d} \leq 1$ and $c_i \geq c_{\min}$. Note that terms $\alpha_{i,t\to\tau+d}$ and $\alpha_{i,t\to\tau} - \frac{1}{c_i}$ above are the updated projected available capacities after the assignment of job j to server i in FLB. Therefore, we show the induction hypothesis after FLB determines the assignment for job j, which finishes the proof of the induction. □

The above key lemma provides an upper bound on the difference $\alpha_{i,t\to t+d} - \alpha_{i,t\to t}$ in terms of the parameters (β, η) of FLB, which is maintained throughout the execution of the algorithm. To understand how this upper bound relates to the capacity feasibility, consider the time t_j at which job j arrives, and let $\alpha_{i,t_j\to\tau}$ denote the projected available capacity after FLB finishes processing job j. Recall that $\alpha_{i,t_j\to\tau}$ is weakly increasing in τ and reaches 1 no later than $\tau = t_j + D$, since all current jobs will certainly be finished by time $t_j + D$. Combining this observation with Lemma 4.4, we obtain the following corollary.

Corollary 4.6 (Capacity Feasibility Condition) *Consider a hypothetical scenario in which* FLB *can assign a job to a server that has no available capacity (which could result in a negative projected capacity level). In the integer-duration environment with inspection-frequency scalar $\gamma = 1$, the current (normalized) available capacity under* FLB *satisfies the following property:*

$$\forall i \in [n], j \in [m]: \quad \alpha_{i,t_j\to t_j} \geq 1 + \frac{\ln\left(\prod_{k\in[D]}\left(1 - \frac{R}{k(R+\eta)}\right) - \frac{(R+\eta)\ln(\beta)}{Rc_{\min}}\right)}{\ln(\beta)}$$

Proof of Proposition 4.2. The capacity feasibility is ensured if $\alpha_{i,t_j\to t_j}$ remains non-negative for all jobs. Thus, a sufficient condition is that the lower bound on this quantity given in Corollary 4.6 is non-negative, which is equivalent to the condition in Proposition 4.2. □

4.2. Primal-Dual Analysis of the Competitive Ratio

In this section, we aim to prove Proposition 4.3. To this end, we first introduce a new linear programming relaxation of the optimal offline benchmark, which we refer to as the *configuration LP* ($\mathcal{P}_{\text{OPT}}$). This LP provides a more refined relaxation than the standard LP, as it (fractionally) assigns each unit of server capacity to a feasible schedule of jobs that can run on this unit. We then conduct a primal-dual analysis based on this configuration LP, resulting in a closed-form characterization of an upper bound on the competitive ratio of FLB as a function of its parameters (η, β).

Configuration LP. Any feasible assignment of jobs to servers can be represented by explicitly specifying a feasible *configuration* for each capacity unit of each server $i \in [n]$. A feasible configuration for server i is defined as a subset of non-overlapping jobs that can be assigned to run on a single capacity unit of this server.

Formally, a subset $S \subseteq [m]$ is a feasible configuration for server i if $(i,j) \in E$ for every job $j \in S$, and for every pair of distinct jobs $j, j' \in S$, it holds that

$$[t_{j'}, t_{j'} + d_{ij'}] \cap [t_j, t_j + d_{ij}] = \emptyset.$$

We denote by $\mathcal{S}_i \subseteq 2^{[m]}$ the set of all feasible configurations for server $i \in [n]$. We now introduce the configuration LP and its dual program as follows:

$$\begin{aligned}
&\underline{\textit{Primal}}: \quad && \max_{\boldsymbol{x} \geq \boldsymbol{0}} \sum\nolimits_{i\in[n]} \sum\nolimits_{S\in\mathcal{S}_i} \sum\nolimits_{j\in S} r_{ij} d_{ij} \cdot x(i,S) \quad \text{s.t.} \\
& && \sum\nolimits_{i\in[n]} \sum\nolimits_{S\in\mathcal{S}_i : j\in S} x(i,S) \leq 1 \qquad j \in [m] \\
& && \sum\nolimits_{S\in\mathcal{S}_i} x(i,S) \leq c_i \qquad i \in [n]
\end{aligned} \tag{$\mathcal{P}_{\text{OPT}}$}$$

$$\begin{aligned}
&\underline{\textit{Dual}}: \quad && \min_{\boldsymbol{\lambda},\boldsymbol{\theta} \geq \boldsymbol{0}} \sum\nolimits_{j\in[m]} \lambda(j) + \sum\nolimits_{i\in[n]} c_i \theta(i) \quad \text{s.t.} \\
& && \sum\nolimits_{j\in S} \lambda(j) + \theta(i) \geq \sum\nolimits_{j\in S} r_{ij} d_{ij} \qquad i \in [n], S \in \mathcal{S}_i
\end{aligned}$$

Here, the decision variable $x(i,S)$ can be interpreted as the probability of assigning a (feasible) configuration S to a single unit of server i. The first set of constraints addresses feasibility from the job perspective: for each job $j \in [m]$, the total assignment probability across all feasible configurations containing job j, summed over all units of all servers, must not exceed 1. The second set of constraints addresses feasibility from the server perspective: for each server $i \in [n]$, the sum of assignment probabilities across all feasible configurations assigned to its units must not exceed its total capacity c_i. The following lemma establishes that this linear program provides a relaxation for the optimal offline benchmark. We defer its formal proof to Section EC.3.2.

Lemma 4.7 *For any sequence of jobs with types $\{r_{ij}, d_{ij}\}_{(i,j)\in E}$, the total reward of the optimal offline benchmark is upper bounded by the optimal objective value of the linear program $\mathcal{P}_{\text{OPT}}$.*

We emphasize that prior related works (e.g., Buchbinder and Naor, 2009; Golrezaei et al., 2014; Feng et al., 2019; Huo and Cheung, 2022) have employed a standard LP formulation for primal-dual analysis, in which decision variables represent the probability of assigning each individual job (rather than configurations) to each server. While this standard LP suffices for obtaining optimal competitive ratios in certain special cases (e.g., $R = D = 1$), it becomes technically challenging—and possibly even theoretically infeasible—to achieve optimal competitive ratios with precise logarithmic dependency on RD (particularly, obtaining the exact constant before $\ln(RD)$) in the general setting (cf., Huo and Cheung, 2022). To circumvent this difficulty, we propose and analyze the configuration LP $\mathcal{P}_{\text{OPT}}$, which allows us to obtain tight, optimal bounds on the competitive ratio.

Proof of Proposition 4.3. We upper bound the competitive ratio of FLB (with inspection-frequency scalar $\gamma = 1$) by constructing a dual solution based on its job assignments as follows.

Initially, set $\lambda(j) \leftarrow 0$ and $\theta(i) \leftarrow 0$ for each job $j \in [m]$ and server $i \in [n]$. Now consider all assignments made by FLB. For each job $j \in [m]$, if FLB assigns job j to server i, update the dual variables as follows:

$$\lambda(j) \leftarrow r_{ij} d_{ij} - \sum_{\tau \in \mathcal{T}_{ij}} \Psi\left(\alpha_{i,t_j \to \tau}\right), \quad \theta(i) \leftarrow \theta(i) + \sum_{\tau \in \mathcal{T}_{ij}} \left(\Psi\left(\alpha_{i,t_j \to \tau} - \frac{1}{c_i}\right) - \Psi\left(\alpha_{i,t_j \to \tau}\right)\right),$$

where $\mathcal{T}_{ij}$ is shorthand for the inspection-time subset $\mathcal{T}_{ij}(\gamma)$ with $\gamma = 1$. Note that the above dual solution is well defined under the assumption that FLB is capacity feasible, meaning that the input argument to the penalty function Ψ is non-negative. This holds since when FLB assigns job j to server i, there is at least one available unit of server i. Therefore, we have $\alpha_{i,t_j \to t_j} \geq \frac{1}{c_i}$. Additionally, for every $\tau \in \mathcal{T}_{ij}$, we have $\alpha_{i,t_j \to \tau} \geq \alpha_{i,t_j \to t_j} \geq \frac{1}{c_i}$, as $\alpha_{i,t_j \to \tau}$ is increasing in τ and $\tau \geq t_j$.

The rest of the proof is done in two steps:

[*Step i*] *Comparing objective values in primal and dual.* Here, we show that the total reward of FLB is a Γ-approximation to the objective value of the constructed dual solution, where $\Gamma = \ln(\beta) \cdot \left(1 + \eta\left(1 + \beta\left(\beta^{\frac{1}{c_{\min}}} - 1\right)\right)\right)$, as stated in Proposition 4.3. To establish this, we analyze the increments of the reward in FLB and the objective value of the constructed dual solution resulting from each job assignment decision separately.

Suppose FLB assigns job j, arriving at time t_j, to server i. The increment of the total reward in FLB due to this assignment is $\Delta(\text{PRIMAL}) = r_{ij} d_{ij}$, while the increment of the objective value of the constructed dual solution can be upper bounded as:

$$\begin{aligned}
\Delta(\text{DUAL}) &= r_{ij} d_{ij} - \sum_{\tau \in \mathcal{T}_{ij}} \Psi\left(\alpha_{i,t_j \to \tau}\right) + \sum_{\tau \in \mathcal{T}_{ij}} c_i \left(\Psi\left(\alpha_{i,t_j \to \tau} - \frac{1}{c_i}\right) - \Psi\left(\alpha_{i,t_j \to \tau}\right)\right) \\
&\overset{(a)}{\leq} \ln(\beta)\left(r_{ij} d_{ij} - \sum_{\tau \in \mathcal{T}_{ij}} \Psi\left(\alpha_{i,t_j \to \tau}\right)\right) - \sum_{\tau \in \mathcal{T}_{ij}} c_i \left(\Psi\left(\alpha_{i,t_j \to \tau}\right) - \Psi\left(\alpha_{i,t_j \to \tau} - \frac{1}{c_i}\right)\right) \\
&= r_{ij} d_{ij} \ln(\beta) - \sum_{\tau \in \mathcal{T}_{ij}} \left(\ln(\beta)\Psi\left(\alpha_{i,t_j \to \tau}\right) + c_i\left(\Psi\left(\alpha_{i,t_j \to \tau}\right) - \Psi\left(\alpha_{i,t_j \to \tau} - \frac{1}{c_i}\right)\right)\right) \\
&\overset{(b)}{\leq} r_{ij} d_{ij} \ln(\beta) - \sum_{\tau \in \mathcal{T}_{ij}} \left(\ln(\beta)\Psi\left(\alpha_{i,t_j \to \tau}\right) + \Psi'\left(\alpha_{i,t_j \to \tau} - \frac{1}{c_i}\right)\right) \\
&\overset{(c)}{\leq} r_{ij} d_{ij} \ln(\beta) + \sum_{\tau \in \mathcal{T}_{ij}} \eta \ln(\beta)\left(1 + \beta\left(\beta^{\frac{1}{c_{\min}}} - 1\right)\right) \\
&= r_{ij} d_{ij} \ln(\beta) + |\mathcal{T}_{ij}| \eta \ln(\beta)\left(1 + \beta\left(\beta^{\frac{1}{c_{\min}}} - 1\right)\right) \\
&\overset{(d)}{\leq} r_{ij} d_{ij} \ln(\beta) + r_{ij} d_{ij} \eta \ln(\beta)\left(1 + \beta\left(\beta^{\frac{1}{c_{\min}}} - 1\right)\right) = \Gamma \cdot r_{ij} d_{ij} = \Gamma \cdot \Delta(\text{PRIMAL}).
\end{aligned}$$

In the above derivation, inequality (a) holds since $\ln(\beta) \geq 1$ for all $\beta \geq e$, and the reduced reward of job j for server i, $r_{ij} d_{ij} - \sum_{\tau \in \mathcal{T}_{ij}} \Psi\left(\alpha_{i,t_j \to \tau}\right)$, is positive by the assumption that the FLB algorithm assigns job j to server i. Moreover, inequality (b) holds since the penalty function Ψ is convex, and thus $c_i\left(\Psi\left(\alpha_{i,t_j \to \tau}\right) - \Psi\left(\alpha_{i,t_j \to \tau} - \frac{1}{c_i}\right)\right) \geq \Psi'\left(\alpha_{i,t_j \to \tau} - \frac{1}{c_i}\right)$. Also, inequality (c) follows from algebraic simplifications, as detailed below:

$$\ln(\beta)\Psi\left(\alpha_{i,t_j \to \tau}\right) + \Psi'\left(\alpha_{i,t_j \to \tau} - \frac{1}{c_i}\right)$$

$$= \left(\ln(\beta)\,\Psi\left(\alpha_{i,t_j\to\tau}\right) + \Psi'\left(\alpha_{i,t_j\to\tau}\right)\right) + \left(\Psi'\left(\alpha_{i,t_j\to\tau} - \frac{1}{c_i}\right) - \Psi'\left(\alpha_{i,t_j\to\tau}\right)\right)$$
$$= -\eta\ln(\beta) - \eta\ln(\beta)\,\beta^{\left(1-\frac{\alpha_{i,t_j\to\tau}}{c_i}\right)}\left(\beta^{\frac{1}{c_i}} - 1\right) \geq -\eta\ln(\beta)\left(1 + \beta\left(\beta^{\frac{1}{c_{\min}}} - 1\right)\right)$$

Finally, inequality (d) holds since $|\mathcal{T}_{ij}| = d_{ij}$ and $r_{ij} \geq 1$.

[*Step ii*] *Checking the feasibility of dual.* Now we show that the constructed dual solution is feasible. First, note that clearly $\forall j \in [m] : \lambda(j) \geq 0$ (as the reduced reward of job j on server i is always positive if FLB assigns job j to server i) and $\forall i \in [n] :\ \theta_i \geq 0$. Now fix an arbitrary server $i \in [n]$ and a feasible configuration $S \in S_i$. By the assignment rule of FLB and the construction of our dual solution, it is guaranteed that for every job $j \in S$, the dual variable $\lambda(j)$ is at least as large as the reduced reward of job j for server i, that is $\lambda(j) \geq r_{ij}d_{ij} - \sum_{\tau\in\mathcal{T}_{ij}} \Psi\left(\alpha_{i,t_j\to\tau}\right)$. Therefore, the dual constraint associated with the primal variable $x(i,S)$ is satisfied if the following inequality holds.

$$\theta(i) - \sum_{j\in S}\sum_{\tau\in\mathcal{T}_{ij}} \Psi\left(\alpha_{i,t_j\to\tau}\right) \geq 0\ . \tag{1}$$

We prove this inequality with a "charging argument." At a high level, our goal is to identify values $\{\theta_{(j,\tau)}(i)\}_{j\in S,\tau\in\mathcal{T}_{ij}}$ such that

$$\forall j \in S, \tau \in \mathcal{T}_{ij} :\ \ \theta_{(j,\tau)}(i) - \Psi(\alpha_{i,t_j\to\tau}) \geq 0 \qquad \text{and} \qquad \sum_{j\in S}\sum_{\tau\in\mathcal{T}_{ij}} \theta_{(j,\tau)}(i) \leq \theta(i)\ , \tag{2}$$

which then implies inequality (1). In the following, we provide details of this charging argument.

Let us introduce an auxiliary notation $\mathcal{T}_{iS} \triangleq \{(j,\tau) : j \in S, \tau \in \mathcal{T}_{ij}\}$ for each server $i \in [n]$ and configuration $S \in \mathcal{S}_i$. Recall that for every configuration $S \in \mathcal{S}_i$, the processing-time intervals of any two distinct jobs $j_1, j_2 \in S$ do not overlap, and thus $\mathcal{T}_{ij_1} \cap \mathcal{T}_{ij_2} = \emptyset$. Consequently, for every pair $(j_1,\tau_1),(j_2,\tau_2) \in \mathcal{T}_{iS}$, if $\tau_1 \neq \tau_2$, then $j_1 \neq j_2$. We now identify a one-to-many correspondence $\sigma : \mathcal{T}_{iS} \to [m]\times[T]$ that satisfies the following three properties.

- SEPARATION: For every pair of distinct elements $(j_1,\tau_1),(j_2,\tau_2) \in \mathcal{T}_{iS}$, it holds that $\sigma(j_1,\tau_1) \cap \sigma(j_2,\tau_2) = \emptyset$.
- FEASIBILITY-I: For every $(j,\tau) \in \mathcal{T}_{iS}$, we have $|\sigma(j,\tau)| = c_i - c_i\alpha_{i,t_j\to\tau}$. Moreover, for any two distinct elements $(j_1,\tau_1),(j_2,\tau_2) \in \sigma(j,\tau)$, it holds that $j_1 \neq j_2$.
- FEASIBILITY-II: For every $(j,\tau) \in \mathcal{T}_{iS}$ and every $(j',\tau') \in \sigma(j,\tau)$, we have $j' < j$, $\tau' \leq \tau$, $\tau' \in \mathcal{T}_{ij'}$, and a unit of server i is assigned to job j' with duration $d_{ij'} > \tau - t_{j'}$.

Before demonstrating the existence of such a correspondence σ, we first illustrate how to construct values $\{\theta_{(j,\tau)}(i)\}_{j\in S,\tau\in\mathcal{T}_{ij}}$ that satisfy condition (2) given the correspondence σ. Specifically, for every $j \in S$ and $\tau \in \mathcal{T}_{ij}$, define

$$\theta_{(j,\tau)}(i) \triangleq \sum_{(j',\tau')\in\sigma(j,\tau)} \left(\Psi\left(\alpha_{i,t_{j'}\to\tau'} - \frac{1}{c_i}\right) - \Psi\left(\alpha_{i,t_{j'}\to\tau'}\right)\right)$$

Note that property SEPARATION and property FEASIBILITY-II of the correspondence σ guarantee the second half of condition (2), i.e., $\sum_{j\in S}\sum_{\tau\in\mathcal{T}_{ij}}\theta_{(j,\tau)}(i)\le\theta(i)$. To establish the first half of condition (2), i.e., $\theta_{(j,\tau)}(i)-\Psi(\alpha_{i,t_j\to\tau})\ge 0$, observe that

$$\begin{aligned}\theta_{(j,\tau)}(i)-\Psi(\alpha_{i,t_j\to\tau}) &= \sum\nolimits_{(j',\tau')\in\sigma(j,\tau)}\left(\Psi\left(\alpha_{i,t_{j'}\to\tau'}-\frac{1}{c_i}\right)-\Psi\left(\alpha_{i,t_{j'}\to\tau'}\right)\right)-\Psi(\alpha_{i,t_j\to\tau})\\ &\overset{(a)}{\ge} \sum\nolimits_{(j',\tau')\in\sigma(j,\tau)}\left(\Psi\left(\alpha_{i,t_{j'}\to\tau}-\frac{1}{c_i}\right)-\Psi\left(\alpha_{i,t_{j'}\to\tau}\right)\right)-\Psi(\alpha_{i,t_j\to\tau})\\ &\overset{(b)}{\ge} \sum\nolimits_{\ell\in[c_i-c_i\alpha_{i,t_j\to\tau}]}\left(\Psi\left(\frac{c_i-\ell}{c_i}\right)-\Psi\left(\frac{c_i-\ell+1}{c_i}\right)\right)-\Psi\left(\alpha_{i,t_j\to\tau}\right)\overset{(c)}{=}-\Psi(1)\overset{(d)}{=}0.\end{aligned}$$

In the above derivation, inequality (a) holds since the projected available capacity $\alpha_{i,t_{j'}\to\tau'}$ is weakly increasing in τ', property FEASIBILITY-II of the correspondence σ ensures $\tau'\le\tau$, and the penalty function Ψ is convex. For inequality (b), properties FEASIBILITY-I and FEASIBILITY-II of the correspondence σ imply that there exist $s\triangleq c_i-c_i\alpha_{i,t_j\to\tau}$ jobs $j_1<j_2<\cdots<j_s<j$, each of which is assigned to a different unit of server i and occupies that unit at time τ. This further implies that, for each $\ell\in[s]$, $\alpha_{i,t_{j_\ell}\to\tau}\le\frac{c_i-\ell+1}{c_i}$. Applying the convexity of the penalty function Ψ completes the argument for inequality (b). Finally, equality (c) holds by simplifying algebra (noticing the telescopic summation), and equality (d) follows since $\Psi(1)=0$.

Now we show the existence of the correspondence σ satisfying properties SEPARATION, FEASIBILITY-I, and FEASIBILITY-II through the following explicit construction. Fix an arbitrary $(j,\tau)\in\mathcal{T}_{iS}$. As mentioned earlier, by definition of the projected available capacity $\alpha_{i,t_j\to\tau}$, there exist $s\triangleq c_i-c_i\alpha_{i,t_j\to\tau}$ jobs $j_1<j_2<\cdots<j_s<j$, each of which is assigned to a distinct unit of server i and occupies this unit at time τ. In other words, for every $\ell\in[s]$, we have $t_{j_\ell}+d_{ij_\ell}\ge\tau$. Moreover, there exists a unique $\tau_\ell\in\mathcal{T}_{ij_\ell}$ such that $\tau_\ell\le\tau<\tau_\ell+1$. By defining

$$\sigma(j,\tau)\triangleq\{(j_\ell,\tau_\ell)\}_{\ell\in[s]},$$

properties FEASIBILITY-I and FEASIBILITY-II are straightforwardly satisfied.

To show property SEPARATION, fix two distinct elements $(j',\tau'),(j'',\tau'')\in\mathcal{T}_{iS}$ such that $j'\le j''$. It suffices to show that $|\tau''-\tau'|\ge 1$. We consider two cases separately. If $j'=j''$, then for $\tau'\ne\tau''$, the construction of inspection-time subset $\mathcal{T}_{ij'}$ directly implies that $|\tau''-\tau'|\ge 1$, as desired. If $j'<j''$, then since both $j',j''\in S$, the inspection-time subsets $\mathcal{T}_{i,j'}$ and $\mathcal{T}_{i,j''}$ do not overlap, meaning that $t_{j'}+d_{ij'}\le t_{j''}$. Thus, we have $\tau'\le\max\mathcal{T}_{ij'}\le t_{j'}+d_{ij'}\le t_{j''}=\min\mathcal{T}_{ij''}\le\tau''$. Moreover, since all durations are integer-valued, we have $\max\mathcal{T}_{ij'}=t_{j'}+d_{ij'}-1$. Hence, $|\tau''-\tau'|=\tau''-\tau'\ge t_{j''}-\max\mathcal{T}_{ij'}\ge 1$, as desired.

Finally, since the total reward of FLB is a Γ-approximation of the objective value of the constructed dual solution (Step i), and the constructed dual solution is feasible (Step ii), invoking the weak duality of the linear program concludes the proof. □

4.3. Putting All the Pieces Together: Proof of Theorem 4.1

Proposition 4.2 and Proposition 4.3 allow us to formulate the best competitive ratio upper bound obtained from the primal-dual analysis of FLB (with inspection-frequency scalar $\gamma = 1$) as the following optimization problem over the penalty parameters (η, β), denoted by $\mathcal{P}_{\text{FLB-INT}}[R, D, c_{\min}]$:

$$\begin{aligned} \min_{\eta,\beta} \quad & \ln(\beta)\cdot\left(1+\eta\left(1+\beta\left(\beta^{\frac{1}{c_{\min}}}-1\right)\right)\right) \qquad \text{s.t.} \\ & \ln(\beta) \geq -\ln\left(\prod_{k\in[D]}\left(1-\frac{R}{k(R+\eta)}\right)-\frac{(R+\eta)\ln(\beta)}{Rc_{\min}}\right) \qquad (\mathcal{P}_{\text{FLB-INT}}[R,D,c_{\min}]) \\ & \eta > 0, \beta \geq e \end{aligned}$$

Here, the first constraint follows from the capacity feasibility condition stated in Proposition 4.2, while the objective and the second constraint follow from Proposition 4.3.

In practice, the decision-maker (e.g., a cloud computing platform) can select the parameters of FLB based on the platform's maximum reward R, maximum duration D, and minimum capacity $c_{\min}$, by numerically solving $\mathcal{P}_{\text{FLB-INT}}[R, D, c_{\min}]$ in the integer-duration environment. To prove the asymptotically optimal competitive ratio stated in Theorem 4.1, we provide the following simple analytical analysis of $\mathcal{P}_{\text{FLB-INT}}[R, D, c_{\min}]$ when $c_{\min}$ goes to ∞. The analysis for $c_{\min} < \infty$ can be found in Section EC.6.

Proof of Theorem 4.1. We focus on the large capacity regime case where $c_{\min} \to \infty$ in $\mathcal{P}_{\text{FLB-INT}}[R, D, c_{\min}]$.[8] We consider two cases based on the magnitude of $R \vee D$.

[*Case i*] Suppose $\ln(R \vee D) \geq (e-1)$. In this case, we assign $\eta = \frac{1}{\ln(R\vee D)}$. We set β such that the first constraint in program $\mathcal{P}_{\text{FLB-INT}}[R, D, \infty]$ binds, i.e.,

$$\ln(\beta) = -\ln\left(\prod_{k\in[D]}\left(1-\frac{R\ln(R\vee D)}{k(R\ln(R\vee D)+1)}\right)\right)$$

It can be verified that constraint $\beta \geq e$ from $\mathcal{P}_{\text{FLB-INT}}[R, D, \infty]$ is satisfied. Therefore, the competitive ratio of FLB (with the aforementioned parameter values) is at most

$$\begin{aligned} & -\left(1+\frac{1}{\ln(R\vee D)}\right)\ln\left(\prod_{k\in[D]}\left(1-\frac{R\ln(R\vee D)}{k(R\ln(R\vee D)+1)}\right)\right) \\ = & -\left(1+\frac{1}{\ln(R\vee D)}\right)\sum_{k\in[D]}\ln\left(1-\frac{R\ln(R\vee D)}{k(R\ln(R\vee D)+1)}\right) = \ln(RD)+\ln\ln(R\vee D)+O(1) \end{aligned}$$

where the last equality holds since

$$-\ln\left(1-\frac{R\ln(R\vee D)}{R\ln(R\vee D)+1}\right) = \ln(R)+\ln\ln(R\vee D)+O(1)$$

and

$$-\sum_{k\in[2:D]}\ln\left(1-\frac{R\ln(R\vee D)}{kR\ln(R\vee D)+1}\right) \leq \sum_{k\in[2:D]}\ln\left(\frac{k}{k-1}\right) = \ln(D)$$

[*Case ii*] Suppose $\ln(R \vee D) \leq e-1$. In this case, we assign $\eta = \frac{R}{e-1}$ and $\beta = e$. It can be verified that the first constraint in $\mathcal{P}_{\text{FLB-INT}}[R, D, \infty]$ is satisfied. Therefore the competitive ratio of FLB (with the aforementioned parameter values) is at most $1 + \frac{R}{e-1} = \Theta(1)$. □

[8] Due to continuity, examining the case where $c_{\min} = \infty$ yields the asymptotic competitive ratio as $c_{\min} \to \infty$. For further details, refer to section EC.6.

Improved competitive ratios for special cases. We can also derive a better bound on the competitive ratio for instances with heterogeneous durations but homogeneous rewards (i.e., $D \geq 1$ but $R = 1$). This special case is motivated by practical considerations and, from a theoretical standpoint, it provides insights into the impact of duration heterogeneity in our model.[9] A similar result for (real-valued) durations is also given in Proposition EC.2.2.

Proposition 4.8 (Competitive Ratio for Integer-valued Durations and Homogeneous Rewards) *For instances with homogeneous rewards (i.e., $R = 1$) in the integer-duration environment, there exists a choice of parameters (η^*, β^*) such that the asymptotic competitive ratio of* FLB*, with inspection-frequency scalar $\gamma^* = 1$ and penalty parameters (η^*, β^*), is at most $H(D) + 2$, where $H(D) \triangleq \sum_{i\in[D]} \frac{1}{i}$ is the D-th harmonic number. Moreover, there exist instances involving a single server with $R = 1, D \geq 1$, for which the asymptotic competitive ratio of any online algorithm (possibly fractional or randomized) against the optimal offline benchmark is at least $H(D)$.*

Proof of Proposition 4.8. We first show the competitive ratio upper bound. Consider instances where all compatible job server pairs (i, j) have the same per-period reward, normalized to $r_{ij} = 1$. In the large capacity regime, program $\mathcal{P}_{\text{FLB-INT}}[1, D, \infty]$ is simplified as

$$\min_{\eta\geq 0,\beta\geq e} \ \ln(\beta)\cdot(1+\eta) \quad \text{s.t.} \quad \ln(\beta) \geq -\ln\left(\prod_{k\in[D]}\left(1-\frac{1}{k(1+\eta)}\right)\right)$$

Is is evident that the first inequality binds in the optimal solution, and therefore, the competitive ratio of FLB can be expressed as $-\ln\left(\prod_{k\in[D]}\left(1-\frac{1}{k(1+\eta)}\right)\right)\cdot(1+\eta)$ as a function of η, which is decreasing in η. Combining with the constraint that $\beta \geq e$, we obtain an upper bound for the competitive ratio of the FLB is as $1 + \eta^{(D)}$ where $\eta^{(D)}$ is the solution of

$$\prod_{k\in[D]}\left(1-\frac{1}{k(1+\eta)}\right) = \frac{1}{e}$$

In the special case of $D = 1$, we obtain $\eta^{(1)} = \frac{1}{e-1}$, $\beta = e$ and FLB recovers the optimal competitive ratio of $\frac{e}{e-1}$ (Feng et al., 2019; Goyal et al., 2025). For $D \geq 2$, we can upper bound $\eta^{(D)}$ with $H(D)$ + 2 and thus obtain a competitive ratio of $H(D) + 2$ with Lemma 4.9 (see its proof in Section EC.3.3).

Next, we construct an adversarial instance involving a single server with capacity c, for which no online algorithm achieves a competitive ratio smaller than $H(D)$. Let $M > c$. Consider an arrival pattern where jobs arrive in D batches, each of size M, at times $0, 0+\epsilon, 0+2\epsilon, \ldots$, for an infinitesimally small ϵ. All jobs in batch $d \in [D]$ have duration d. We assume these batches arrive sequentially in the order $1, 2, 3, \ldots$ (thus, all jobs in batch $i > j$ arrive after jobs in batch j), and the index of the last batch is chosen adversarially. For any given online algorithm $\mathcal{A}$, let x_i denote the (fractional) number of jobs from batch i accepted by $\mathcal{A}$. Note that

[9] The impact of the reward heterogeneity for online resource allocation has been studied in the literature (e.g., Ball and Queyranne, 2009; Feldman et al., 2009; Ma and Simchi-Levi, 2020; Ekbatani et al., 2023) for non-reusable resources.

if the last batch is d, the optimal offline benchmark assigns all c units exclusively to jobs in batch d (and none to jobs from batches $i \in [d-1]$). Therefore, if $\mathcal{A}$ is $(H(D)-\delta)$-competitive for some $\delta > 0$, it follows that for all durations $d \in [D]$,

$$\sum\nolimits_{i\in[d]} i x_i > \frac{c \cdot d}{H(D)}.$$

Multiplying the inequality corresponding to d by $(\frac{1}{d} - \frac{1}{d+1})$ for $d < D$, and by $\frac{1}{D}$ for $d = D$, and summing all terms, we obtain $\sum_{i\in[D]} x_i > c$. This inequality contradicts the capacity feasibility of $\mathcal{A}$ when all batches arrive. □

Lemma 4.9 *Let λ_D be the solution to the equation $e^{-1} = \prod_{k\in[D]} \left(1 - \frac{\lambda_D}{k}\right)$. Then $\frac{1}{\lambda_D} \le H(D) + 2$.*

Remark 4.1 (Competitive Ratio for Homogeneous Rewards and Durations) *In the special case of $R = D = 1$, the competitive ratio of the* FLB *with $\gamma = 1, \eta = \frac{1}{e-1}, \beta = e$ is $\frac{e}{e-1}$. Moreover, competitive ratio of $\frac{e}{e-1}$ is the optimal among all (possibly fractional and randomized) online algorithms (Feng et al., 2019; Goyal et al., 2025).*

4.4. Proof Sketch of Theorem 3.1 for Real-valued Durations

In this section, we explain how to extend our previous analysis—which assumed integer-valued durations—to our main setting with real-valued durations. The formal analysis of Theorem 3.1 can be found in Section EC.2. The high-level approach resembles the one used for integer-valued durations: we first identify a sufficient condition for the capacity feasibility in Proposition 4.10. Next we characterize its competitive ratio in Proposition 4.11. Finally, we formulate the problem of identifying the asymptotically optimal competitive ratio as an optimization program $\mathcal{P}_{\text{FLB-REAL}}[R, D, c_{\min}]$ induced by the propositions in the first two steps. One key difference is that the inspection frequency scalar γ is no longer fixed at one. In fact, the asymptotically optimal competitive ratio is achieved when $\gamma \ge 2$.

Proposition 4.10 (Capacity Feasibility of FLB) FLB *is capacity feasible if integer-valued inspection-frequency scalar $\gamma \in \mathbb{N}$ and penalty parameters $(\eta, \beta) \in \mathbb{R}^2_+$ satisfy*

$$\ln(\beta) \ge -\ln\left(\prod\nolimits_{k\in[\lceil\gamma D\rceil]} \left(1 - \frac{R}{k(R+\gamma\eta)}\right)\right) - \ln\left(1 + \left(\gamma + \frac{R}{\eta}\right)\left(1 - \frac{(\eta\gamma + R)\ln(\beta)}{R c_{\min}}\right) - \frac{R}{\eta}\left(1 + \frac{\eta}{R+\gamma\eta}\right)^{\gamma}\right)$$
$$+ \ln\left(\frac{(\gamma+1)(R+\gamma\eta)}{\gamma\eta}\right) + \ln\left(\prod\nolimits_{k\in[\gamma+1]}\left(1 - \frac{R}{k(R+\gamma\eta)}\right)\right)$$

The proof of Proposition 4.10 extends the induction argument used in Proposition 4.2. We maintain a similar invariant for projected available capacities, as formalized in Lemma EC.2.1. The key distinction arises from considering general $\gamma \in \mathbb{N}$ in FLB. This requires analyzing the projected available capacities difference $\alpha_{i,t\to\tau+d} - \alpha_{i,t\to\tau}$ for $d \in \mathbb{R}_+$, which yields different closed-form expressions for $d \le 1$ and $d \ge 1$. In contrast, the integer-valued duration case with $\gamma = 1$ only needs to consider $d \in \mathbb{N}$, eliminating the need for $d \le 1$ case.

Proposition 4.11 (Competitive Ratio of Capacity Feasible FLB) *For every $\gamma \geq 2$, $\eta > 0$ and $\beta \geq e$, the competitive ratio of a capacity feasible* FLB *is at most*

$$\frac{\gamma}{\gamma-1} \cdot \ln(\beta) \cdot \left(1 + \gamma\eta \left(1 + \beta \left(\beta^{\frac{1}{c_{\min}}} - 1\right)\right)\right)$$

The proof of Proposition 4.11 follows a similar primal-dual analysis with the configuration LP $\mathcal{P}_{\text{OPT}}$. The dual solution construction mirrors the one in the proof of Proposition 4.2. The argument which compares the objective values in the primal and dual is also similar. The key difference lies in the feasibility of the dual constraint. Unlike for the integer-valued durations, the same dual solution construction only guarantees *approximate* dual feasibility, with a multiplicative factor of $\gamma/(\gamma-1)$.[10] To establish this approximate feasibility, we develop a more involved charging argument through an extended construction.

5. Tightness of the Competitive Ratio

In this section, we demonstrate that the competitive ratio of FLB is asymptotically optimal among all online algorithms, including those that allow randomization or fractional assignments. We construct an adversarial problem instance (illustrated in Figure 2) with a single server. Jobs arrive sequentially with the shortest jobs with the lowest rewards arriving first. As time progresses, both job duration and reward steadily increase until arrivals abruptly stop at an adversarial time. Intuitively speaking, since the algorithm does not know when the arrivals will end, it must hedge against such an abrupt end of arrival by accepting a minimum number of arriving jobs to maintain the desirable competitive ratio. This necessity limits its ability to accept the (ex-post) most valuable jobs.

We now formalize this argument by restating and proving the lower bound result in Theorem 3.1.

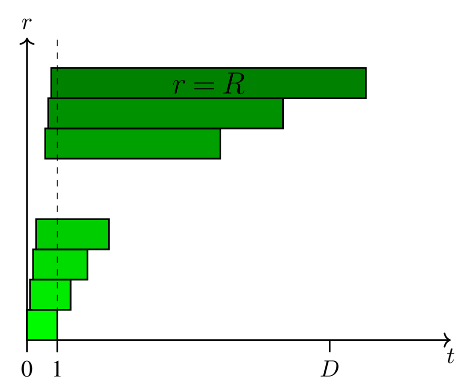

Figure 2 The graphical illustration of the worst-case instance in Proposition 5.1. The darker green means a job with higher reward.

Proposition 5.1 (Restating Negative Result in Theorem 3.1) *For instances involving a single server with $R \geq 1$ and $D \geq 1$, the competitive ratio of any online algorithm (possibly fractional or randomized) against the optimal offline benchmark is at least* $\ln(RD) + \Omega(1)$.

[10] This factor is unavoidable in our dual construction, and thus we cannot use FLB with inspection frequency scalar $\gamma = 1$ for real-valued durations.

Proof. By Yao's lemma (Yao, 1977), it suffices to construct a distribution over problem instances and argue no fractional deterministic online algorithm achieves a competitive ratio better than $\ln(RD)$.

Let $M \in \mathbb{N}$ be a sufficiently large integer. We construct a distribution over M instances $\{I_k\}_{k\in[M]}$ as follows. In all M instances, there are $m \triangleq M$ jobs and a single server with initial capacity $c \triangleq 1$.[11] Jobs of instance I_k are constructed as follows: every job $j \in [k]$ has reward $r_j^{(k)} = R^{j/M}$ and duration $d_j^{(k)} = \lfloor D^{j/M} \rfloor$, arriving at time j/M. Each job $j \in [k+1 : M]$ arrives at time j/M but is incompatible with the server.

We set the probability p_M of instance I_M as $p_M \triangleq \frac{r_1^{(1)} d_1^{(1)}}{r_M^{(M)} d_M^{(M)}}$, and the probability p_k of instance I_k as $p_k \triangleq \frac{r_1^{(1)} d_1^{(1)}}{r_k^{(k)} d_k^{(k)}} - \frac{r_1^{(1)} d_1^{(1)}}{r_{k+1}^{(k+1)} d_{k+1}^{(k+1)}}$ for every $k \in [M-1]$. Under our construction, the expected total reward of both the optimal offline benchmark and the configuration LP benchmark $\mathcal{P}_{\text{OPT}}$ is $\sum_{k\in[M]} p_k r_k^{(k)} d_k^{(k)} = M\left(1 - \frac{1}{(RD)^{1/M}}\right) + \Omega(1)$, where the $\Omega(1)$ term arises due to rounding in the duration construction. On the other hand, for an arbitrary deterministic online algorithm ALG, it cannot distinguish among instances $\{I_k\}_{k=j}^{M}$ upon the arrival of job j. Thus, the expected total reward of ALG can be upper bounded by

$$\max_{\mathbf{y}\geq \mathbf{0}} \sum_{k\in[M]} \left(\sum_{\tau\in[k:M]} p_\tau\right) r_k^{(k)} d_k^{(k)} y(k) \quad \text{s.t.} \quad \sum_{k\in[M]} y(k) \leq 1,$$

where variable $y(k)$ denotes the fraction of the server assigned to job k by ALG. Solving this simple LP, the total reward of ALG is at most $\max_{k\in[M]} \left(\sum_{\tau\in[k:M]} p_\tau\right) r_k^{(k)} d_k^{(k)} = 1$. Therefore, the optimal competitive ratio is at least $M\left(1 - \frac{1}{(RD)^{1/M}}\right) + \Omega(1)$. By letting M approach infinity, we obtain $\lim_{M\to\infty} M\left(1 - \frac{1}{(RD)^{1/M}}\right) + \Omega(1) = \ln(RD) + \Omega(1)$, as desired. □

6. Numerical Experiments

In this section, we compare the numerical performance of FLB against several benchmarks:

- <u>OPT:</u> The optimal offline benchmark for the given instance.
- <u>GREEDY:</u> A greedy algorithm that assigns each arriving job j to the compatible server with available capacity that offers the maximum total reward $r_{ij} d_{ij}$.
- <u>BALANCE:</u> The BALANCE algorithm, which assigns each arriving job j to the server with the highest positive reduced reward $r_{ij} d_{ij} - \Psi_{\text{B}}\left(\alpha_{i,t_j \to t_j}\right)$ (if any), where $\Psi_{\text{B}}(x) = \frac{RD}{e-1}\left(e^{1-x} - 1\right)$. Note that to maintain capacity feasibility, it must satisfy $\Psi_{\text{B}}(0) \geq RD$.[12]

We consider two types of instances for our numerical experiments. First, in Section 6.1, we evaluate the performance of our algorithm and the benchmarks on the worst-case instance presented in Section 5, numerically demonstrating that (i) the benchmarks perform arbitrarily poorly on this instance, and (ii) FLB achieves the worst-case theoretical guarantee outlined earlier in the paper. The results highlight how FLB mitigates worst-case performance by strategically reserving sufficient capacity for potential high-duration, high-reward jobs in the future. Second, in Section 6.2, we evaluate performance on more practical, beyond worst-case

[11] By duplicating jobs, our proof can also be extended to prove the same negative result for large capacities.

[12] Golrezaei et al. (2014) define $\Psi_{\text{B}}(x) = \frac{e^{1-x}-1}{e-1}$, which matches our definition when $R = D = 1$.

instances, and investigate how "job congestion" affects the results. Our findings indicate that, although FLB offers limited benefits in fully utilized or under-utilized scenarios, it has clear advantages in intermediate utilization regimes.

6.1. Against the Worst-Case Instance

Setting. Inspired by the worst-case instance introduced in Section 5, we consider a family of $M = 1000$ instances, each with a single server ($n = 1$), capacity $c = 200$, and parameters $R = D = 10$. In instance $m \in [1 : M]$, there are m jobs arriving within the interval $[0, 1]$, with durations and rewards increasing exponentially over time according to a geometric progression: each job $j \in [1 : m]$ arrives at time $t_j = \frac{j-1}{1000}$ and has reward $r_j = R^{t_j}$ and duration $d_j = \lfloor D^{t_j} \rfloor$. Importantly, we require our algorithm and benchmarks to provide robust performance guarantees against all instances in this family, as though these $M = 1000$ jobs could continue arriving indefinitely, yet the adversary may stop arrivals at any point (see Figure 2).

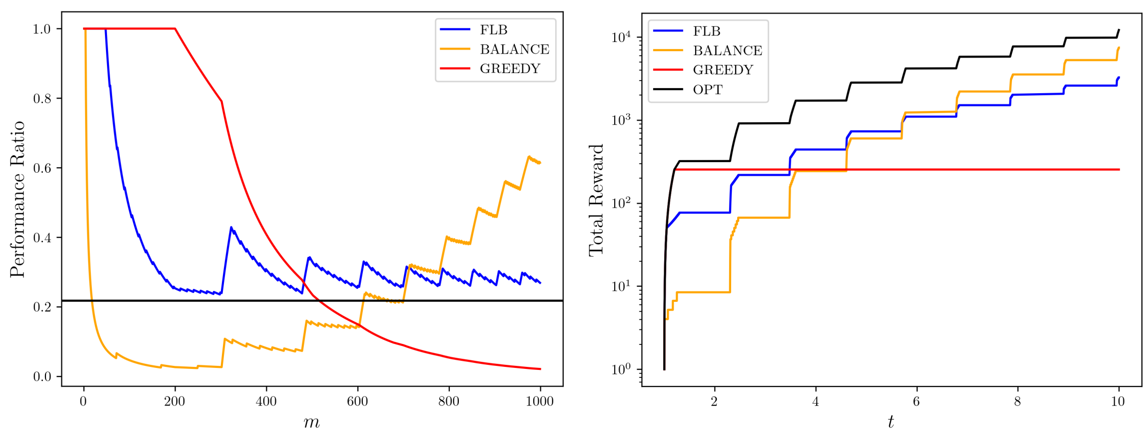

Figure 3 The left panel illustrates the performance ratio of the algorithms relative to the optimal offline solution, with the solid black line indicating the tight theoretical guarantee $y = \frac{1}{\ln(RD)} \approx 0.217$. The right panel presents the total reward for the instance with $m = 1000$. The observed jumps result from the discrete nature of durations, which take integer values.

Discussion. The performance of the algorithms compared to the optimal offline solution is shown in Figure 3. As observed, GREEDY accepts all jobs until reaching full capacity at job number 200, after which it loses competitiveness. In contrast, BALANCE initially assigns very conservatively due to the high penalty required for feasibility, resulting in a poor performance ratio if the problem instance has fewer than approximately 500 jobs. However, FLB consistently maintains its performance ratio above the theoretical bound $\frac{1}{\ln(RD)}$, as established theoretically in Section 4.

6.2. Against a Randomized Practical Instance

Setting. We consider $n = 3$ servers, each with capacity $c_1 = c_2 = c_3 \in \{5, 10, 20, 50\}$. A total of $m = 500$ jobs arrive following a Poisson arrival process with rate λ. For each $\lambda \in [1 : 250]$, we evaluate the algorithms on 100 different random instances. The rewards are drawn independently and identically from a truncated normal distribution (bounded within $[0, 10]$) with mean $\mu = 2$ and standard deviation $\sigma = 3$. Job durations are integer-valued, obtained by taking the ceiling of random variables sampled from the same truncated normal distribution. In Figure 4, we report confidence intervals for the mean performance of each algorithm across different values of $\lambda \in [1:250]$ for each $c_i \in \{5, 10, 20, 50\}$. Additionally, in Figure 5, we present box-and-whisker diagrams illustrating the performance of the algorithms across 100 different random instances for $(c, \lambda) \in \{(10, 50), (50, 100)\}$.

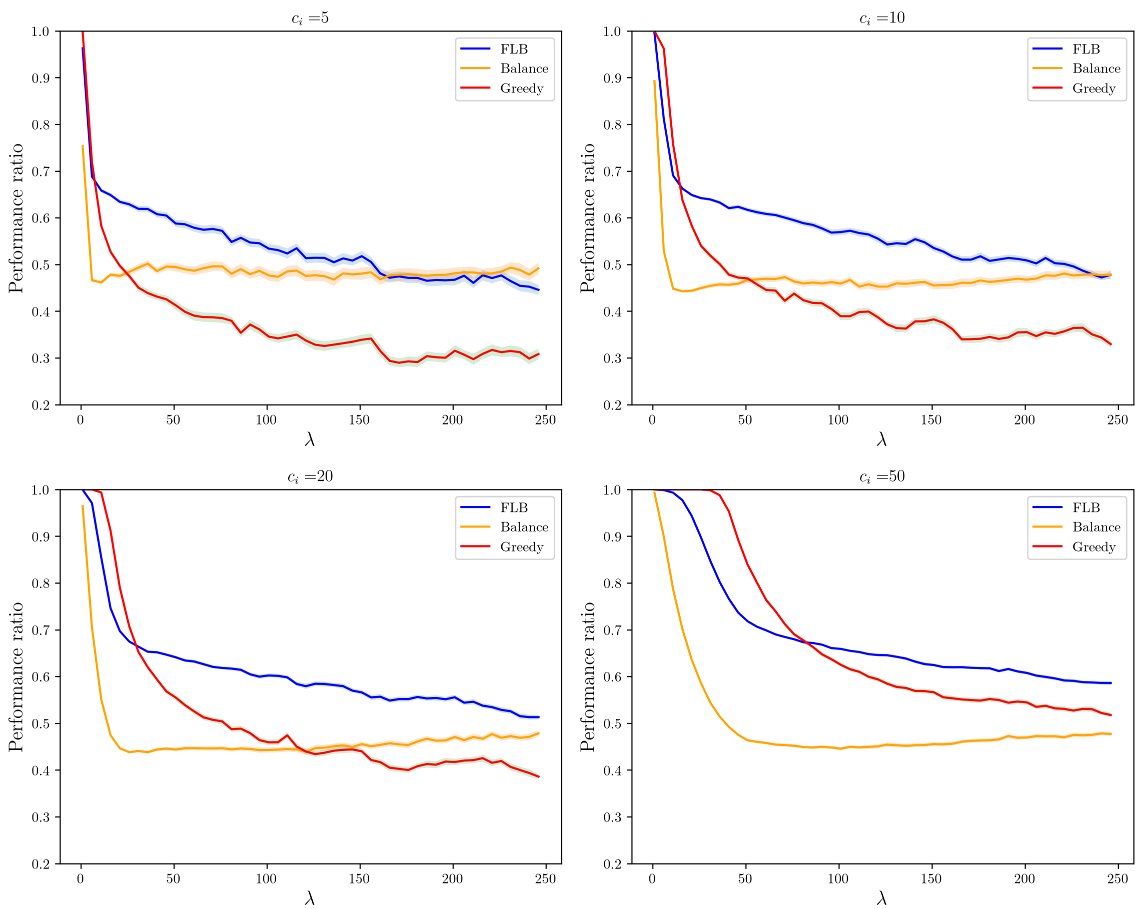

Figure 4 Performance comparison and confidence intervals for FLB and alternative algorithms on the random instance for different values of c_i. Parameters: $n = 3, R = D = 10, m = 500, c_1 = c_2 = c_3$

Discussion. For small values of λ and high capacity c, capacity management is less critical since the system is underutilized. As a result, GREEDY outperforms both BALANCE and FLB. Additionally, in these regimes, BALANCE rejects a significant number of jobs due to its conservative penalty structure. Conversely, for large λ, the majority of jobs arrive before $t = 1$, reducing FLB's advantage in managing capacity for future arrivals. Consequently, the performance gap between FLB and BALANCE diminishes as λ grows. In the intermediate

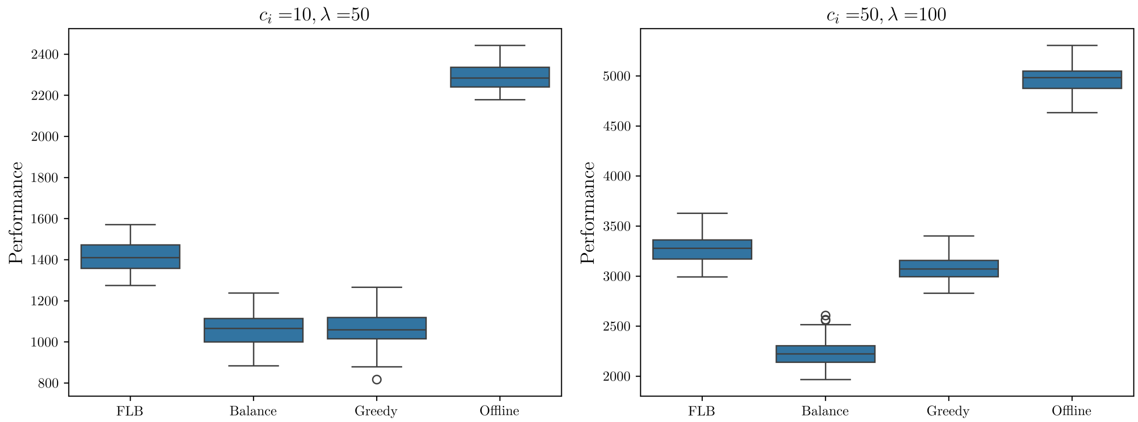

Figure 5 Box-and-whisker diagram for FLB and alternative algorithms on the 100 random instance for values of $(c, \lambda) \in \{(10,50),(50,100)\}$. Parameters: $n = 3, R = D = 10, m = 500, c_1 = c_2 = c_3$

regime, where job arrivals are neither too high to cause most jobs to arrive before $t = 1$ nor too low to eliminate the need for capacity management, FLB significantly outperforms the other algorithms.

7. Conclusion & Future Directions

Motivated by applications in cloud computing marketplaces, we introduced the online job assignment problem. We proposed Forward-Looking BALANCE (FLB) and characterized its asymptotic competitive ratio in various settings using a novel primal-dual analysis, obtaining optimal or nearly optimal results. Although closely related to extensively studied models such as online packing/covering and reusable-resource matching, existing algorithms from these literature do not provide strong guarantees due to fundamental features in our model, such as continuous arrivals and heterogeneous job durations and rewards. We developed new techniques to address these challenges.

Future research. Several open questions remain. First, while this work mainly focuses on the large-capacity regime, deriving nearly optimal competitive ratios under small capacity would also be interesting; even for the special case $R = D = 1$, the best known competitive ratio is $0.589 < 1 - 1/e$ in that regime (Delong et al., 2024). Second, in certain practical settings, the platform may allow only discrete sets of possible rewards and job durations (Ma and Simchi-Levi, 2020). Can we apply FLB to obtain more refined competitive-ratio guarantees depending explicitly on these discrete sets? Third, platforms might leverage machine-learning predictions of future job arrivals. Given the forward-looking nature of FLB, could such forecasts be incorporated into the algorithm? As another research question, some platforms allow deferring jobs if all servers are occupied through wait lists, or they allow for advanced booking. It would be interesting to analyze the performance of FLB or develop new algorithms for these variants. Finally, one can consider the variant of the problem with machine-learning-based “advice,” similar to Choo et al. (2025); Mahdian et al. (2007), and study the characterization of the robustness-consistency tradeoff.

References

Ittai Abraham, Moshe Babaioff, Shaddin Dughmi, and Tim Roughgarden. Combinatorial auctions with restricted complements. In Boi Faltings, Kevin Leyton-Brown, and Panos Ipeirotis, editors, *Proceedings of the 13th ACM Conference on Electronic Commerce, EC 2012, Valencia, Spain, June 4-8, 2012*, pages 3–16. ACM, 2012.

Gagan Aggarwal, Gagan Goel, Chinmay Karande, and Aranyak Mehta. Online vertex-weighted bipartite matching and single-bid budgeted allocations. In Dana Randall, editor, *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2011, San Francisco, California, USA, January 23-25, 2011*, pages 1253–1264. SIAM, 2011.

Mohammad Reza Aminian, Will Ma, and Linwei Xin. Real-time personalized order holding. *Available at SSRN 4644495*, 2023.

Ali Aouad and Will Ma. A nonparametric framework for online stochastic matching with correlated arrivals. In Kevin Leyton-Brown, Jason D. Hartline, and Larry Samuelson, editors, *Proceedings of the 24th ACM Conference on Economics and Computation, EC 2023, London, United Kingdom, July 9-12, 2023*, page 114. ACM, 2023.

Baruch Awerbuch, Yossi Azar, and Serge A. Plotkin. Throughput-competitive on-line routing. In *34th Annual Symposium on Foundations of Computer Science, Palo Alto, California, USA, 3-5 November 1993*, pages 32–40. IEEE Computer Society, 1993.

Yossi Azar, Bala Kalyanasundaram, Serge A. Plotkin, Kirk Pruhs, and Orli Waarts. On-line load balancing of temporary tasks. *Journal of Algorithms*, 22(1):93–110, 1997.

Yossi Azar, Inna Kalp-Shaltiel, Brendan Lucier, Ishai Menache, Joseph Naor, and Jonathan Yaniv. Truthful online scheduling with commitments. In Tim Roughgarden, Michal Feldman, and Michael Schwarz, editors, *Proceedings of the Sixteenth ACM Conference on Economics and Computation, EC '15, Portland, OR, USA, June 15-19, 2015*, pages 715–732. ACM, 2015.

Yossi Azar, Niv Buchbinder, T.-H. Hubert Chan, Shahar Chen, Ilan Reuven Cohen, Anupam Gupta, Zhiyi Huang, Ning Kang, Viswanath Nagarajan, Joseph Naor, and Debmalya Panigrahi. Online algorithms for covering and packing problems with convex objectives. In Irit Dinur, editor, *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, pages 148–157. IEEE Computer Society, 2016.

Michael O. Ball and Maurice Queyranne. Toward robust revenue management: Competitive analysis of online booking. *Operations Research*, 57(4):950–963, 2009.

Nikhil Bansal, Alberto Caprara, and Maxim Sviridenko. Improved approximation algorithms for multidimensional bin packing problems. In *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2006), 21-24 October 2006, Berkeley, California, USA, Proceedings*, pages 697–708. IEEE Computer Society, 2006.

Niv Buchbinder and Rica Gonen. Incentive compatible mulit-unit combinatorial auctions: A primal dual approach. *Algorithmica*, 72 (1):167–190, 2015.

Niv Buchbinder and Joseph Naor. Online primal-dual algorithms for covering and packing. *Mathematics of Operations Research*, 34 (2):270–286, 2009.

Niv Buchbinder, Kamal Jain, and Joseph Naor. Online primal-dual algorithms for maximizing ad-auctions revenue. In Lars Arge, Michael Hoffmann, and Emo Welzl, editors, *Algorithms - ESA 2007, 15th Annual European Symposium, Eilat, Israel, October 8-10, 2007, Proceedings*, volume 4698 of *Lecture Notes in Computer Science*, pages 253–264. Springer, 2007.

Ran Canetti and Sandy Irani. Bounding the power of preemption in randomized scheduling. In *Proceedings of the Twenty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '95, page 606–615, New York, NY, USA, 1995. Association for Computing Machinery. ISBN 0897917189.

Shuchi Chawla, Nikhil R. Devanur, Alexander E. Holroyd, Anna R. Karlin, James B. Martin, and Balasubramanian Sivan. Stability of service under time-of-use pricing. In Hamed Hatami, Pierre McKenzie, and Valerie King, editors, *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*, pages 184–197. ACM, 2017.

Shuchi Chawla, J. Benjamin Miller, and Yifeng Teng. Pricing for online resource allocation: Intervals and paths. In Timothy M. Chan, editor, *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*, pages 1962–1981. SIAM, 2019.

Davin Choo, Billy Jin, and Yongho Shin. Learning-augmented online bipartite fractional matching. *arXiv preprint arXiv:2505.19252*, 2025.

Steven Delong, Alireza Farhadi, Rad Niazadeh, Balasubramanian Sivan, and Rajan Udwani. Online bipartite matching with reusable resources. *Mathematics of Operations Research*, 49(3):1825–1854, 2024.

Nikhil R. Devanur, Kamal Jain, Balasubramanian Sivan, and Christopher A. Wilkens. Near optimal online algorithms and fast approximation algorithms for resource allocation problems. *Journal of the ACM (JACM)*, 66(1):7:1–7:41, 2019.

Kurt Eisemann. The trim problem. *Management Science*, 3(3):279–284, 1957. ISSN 00251909, 15265501.

Farbod Ekbatani, Yiding Feng, and Rad Niazadeh. Online resource allocation with buyback: Optimal algorithms via primal-dual. In Kevin Leyton-Brown, Jason D. Hartline, and Larry Samuelson, editors, *Proceedings of the 24th ACM Conference on Economics and Computation, EC 2023, London, United Kingdom, July 9-12, 2023*, page 583. ACM, 2023.

Jon Feldman, Nitish Korula, Vahab S. Mirrokni, S. Muthukrishnan, and Martin Pál. Online ad assignment with free disposal. In Stefano Leonardi, editor, *Internet and Network Economics, 5th International Workshop, WINE 2009, Rome, Italy, December 14-18, 2009. Proceedings*, volume 5929 of *Lecture Notes in Computer Science*, pages 374–385. Springer, 2009.

Yiding Feng and Rad Niazadeh. Batching and optimal multistage bipartite allocations. *Management Science*, 71(5):4108–4130, 2025.

Yiding Feng, Rad Niazadeh, and Amin Saberi. Linear programming based online policies for real-time assortment of reusable resources. *Chicago Booth Research Paper*, 2019.

Yiding Feng, Rad Niazadeh, and Amin Saberi. Robustness of online inventory balancing algorithm to inventory shocks. *Available at SSRN 3795056*, 2021.

Zhichao Feng, Milind Dawande, Ganesh Janakiraman, and Anyan Qi. Online learning and capacity management in cloud-cost optimization. *Available at SSRN 4812792*, 2024.

Negin Golrezaei, Hamid Nazerzadeh, and Paat Rusmevichientong. Real-time optimization of personalized assortments. *Management Science*, 60(6):1532–1551, 2014.

Xiao-Yue Gong, Vineet Goyal, Garud N. Iyengar, David Simchi-Levi, Rajan Udwani, and Shuangyu Wang. Online assortment optimization with reusable resources. *Management Science*, 68(7):4772–4785, 2022.

Vineet Goyal and Rajan Udwani. Online matching with stochastic rewards: Optimal competitive ratio via path-based formulation. *Operations Research*, 71(2):563–580, 2023.

Vineet Goyal, Garud Iyengar, and Rajan Udwani. Asymptotically optimal competitive ratio for online allocation of reusable resources. *Operations Research*, 0(0):0, 2025.

Yu-Ju Hong, Jiachen Xue, and Mithuna Thottethodi. Dynamic server provisioning to minimize cost in an iaas cloud. In Arif Merchant, Kimberly Keeton, and Dan Rubenstein, editors, *SIGMETRICS 2011, Proceedings of the 2011 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, San Jose, CA, USA, 07-11 June 2011 (Co-located with FCRC 2011)*, pages 147–148. ACM, 2011.

Zhiyi Huang and Xinkai Shu. Online stochastic matching, poisson arrivals, and the natural linear program. In Samir Khuller and Virginia Vassilevska Williams, editors, *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021*, pages 682–693. ACM, 2021.

Zhiyi Huang and Qiankun Zhang. Online primal dual meets online matching with stochastic rewards: Configuration LP to the rescue. *SIAM Journal on Computing*, 53(5):1217–1256, 2024.

Zhiyi Huang, Hanrui Jiang, Aocheng Shen, Junkai Song, Zhiang Wu, and Qiankun Zhang. Online matching with stochastic rewards: Advanced analyses using configuration linear programs. In Jugal Garg, Max Klimm, and Yuqing Kong, editors, *Web and Internet Economics - 19th International Conference, WINE 2023, Shanghai, China, December 4-8, 2023, Proceedings*, volume 14413 of *Lecture Notes in Computer Science*, pages 384–401. Springer, 2023.

Zhiyi Huang, Qiankun Zhang, and Yuhao Zhang. Adwords in a panorama. *SIAM Journal on Computing*, 53(3):701–763, 2024.

Tianming Huo and Wang Chi Cheung. Online reusable resource allocations with multi-class arrivals. *Available at SSRN 4320423*, 2022.

Kamal Jain, Mohammad Mahdian, Evangelos Markakis, Amin Saberi, and Vijay V. Vazirani. Greedy facility location algorithms analyzed using dual fitting with factor-revealing LP. *Journal of the ACM (JACM)*, 50(6):795–824, 2003.

Huiwen Jia, Cong Shi, and Siqian Shen. Online learning and pricing for service systems with reusable resources. *Operations Research*, 72(3):1203–1241, 2024.

Bala Kalyanasundaram and Kirk Pruhs. An optimal deterministic algorithm for online b-matching. *Theoretical Computer Science*, 233(1-2):319–325, 2000.

Richard M. Karp, Umesh V. Vazirani, and Vijay V. Vazirani. An optimal algorithm for on-line bipartite matching. In Harriet Ortiz, editor, *Proceedings of the 22nd Annual ACM Symposium on Theory of Computing, May 13-17, 1990, Baltimore, Maryland, USA*, pages 352–358. ACM, 1990.

Ian A. Kash, Peter B. Key, and Warut Suksompong. Simple pricing schemes for the cloud. *ACM Transactions on Economics and Computation (TEAC)*, 7(2):7:1–7:27, 2019.

Richard J. Lipton and Andrew Tomkins. Online interval scheduling. In Daniel Dominic Sleator, editor, *Proceedings of the Fifth Annual ACM-SIAM Symposium on Discrete Algorithms. 23-25 January 1994, Arlington, Virginia, USA*, pages 302–311. ACM/SIAM, 1994.

Brendan Lucier, Ishai Menache, Joseph (Seffi) Naor, and Jonathan Yaniv. Efficient online scheduling for deadline-sensitive jobs: Extended abstract. In *Proceedings of the Twenty-Fifth Annual ACM Symposium on Parallelism in Algorithms and Architectures*, SPAA '13, page 305–314, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450315722.

Will Ma and David Simchi-Levi. Algorithms for online matching, assortment, and pricing with tight weight-dependent competitive ratios. *Operations Research*, 68(6):1787–1803, 2020.

Mohammad Mahdian, Hamid Nazerzadeh, and Amin Saberi. Allocating online advertisement space with unreliable estimates. In *Proceedings of the 8th ACM conference on Electronic commerce*, pages 288–294, 2007.

Vahideh Manshadi, Scott Rodilitz, Daniela Saban, and Akshaya Suresh. Online algorithms for matching platforms with multichannel traffic. *Management Science*, 0(0):0, 2025.

Aranyak Mehta, Amin Saberi, Umesh V. Vazirani, and Vijay V. Vazirani. Adwords and generalized online matching. *Journal of the ACM (JACM)*, 54(5):22, 2007.

Kim Thang Nguyen. Online primal-dual algorithms with configuration linear programs. In Yixin Cao, Siu-Wing Cheng, and Minming Li, editors, *31st International Symposium on Algorithms and Computation, ISAAC 2020, December 14-18, 2020, Hong Kong, China (Virtual Conference)*, volume 181 of *LIPIcs*, pages 45:1–45:16. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020.

Paat Rusmevichientong, Mika Sumida, Huseyin Topaloglu, and Yicheng Bai. Revenue management with heterogeneous resources: Unit resource capacities, advance bookings, and itineraries over time intervals. *Operations Research*, 71(6):2196–2216, 2023.

David Simchi-Levi, Zeyu Zheng, and Feng Zhu. On greedy-like policies in online matching with reusable network resources and decaying rewards. *Management Science*, 0(0):0, 2025.

Rajan Udwani. Adwords with unknown budgets and beyond. *Management Science*, 71(2):1009–1026, 2025.

José Verschae and Andreas Wiese. On the configuration-lp for scheduling on unrelated machines. *Journal of Scheduling*, 17(4): 371–383, 2014.

Changjun Wang, Weidong Ma, Tao Qin, Xujin Chen, Xiaodong Hu, and Tie-Yan Liu. Selling reserved instances in cloud computing. In Qiang Yang and Michael J. Wooldridge, editors, *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 224–231. AAAI Press, 2015.

Wei Wang, Di Niu, Baochun Li, and Ben Liang. Dynamic cloud resource reservation via cloud brokerage. In *IEEE 33rd International Conference on Distributed Computing Systems, ICDCS 2013, 8-11 July, 2013, Philadelphia, Pennsylvania, USA*, pages 400–409. IEEE Computer Society, 2013.

Andrew Chi-Chih Yao. Probabilistic computations: Toward a unified measure of complexity (extended abstract). In *18th Annual Symposium on Foundations of Computer Science, Providence, Rhode Island, USA, 31 October - 1 November 1977*, pages 222–227. IEEE Computer Society, 1977.

Sharrukh Zaman and Daniel Grosu. An online mechanism for dynamic VM provisioning and allocation in clouds. In Rong Chang, editor, *2012 IEEE Fifth International Conference on Cloud Computing, Honolulu, HI, USA, June 24-29, 2012*, pages 253–260. IEEE Computer Society, 2012.

Hong Zhang, Hongbo Jiang, Bo Li, Fangming Liu, Athanasios V. Vasilakos, and Jiangchuan Liu. A framework for truthful online auctions in cloud computing with heterogeneous user demands. *IEEE Transactions on Computers*, 65(3):805–818, 2016a.

Jia Zhang, Weidong Ma, Tao Qin, Xiaoming Sun, and Tie-Yan Liu. Randomized mechanisms for selling reserved instances in cloud computing. In Satinder Singh and Shaul Markovitch, editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 750–757. AAAI Press, 2017.

Rui Zhang, Kui Wu, Minming Li, and Jianping Wang. Online resource scheduling under concave pricing for cloud computing. *IEEE Transactions on Parallel and Distributed Systems*, 27(4):1131–1145, 2016b.

This page is intentionally blank. The e-companion (EC) starts next page.

EC.1. Further Related Work

Here is a summary of further related prior work in the literature directly related to our work.

Job assignment in cloud computing platforms. In the cloud computing setting specifically, prior work uses assumptions such as preemptable jobs Lucier et al. (2013); Canetti and Irani (1995); Aminian et al. (2023) and delayed commitments Lucier et al. (2013); Zaman and Grosu (2012); Azar et al. (2015), neither of which our approach requires. Other work additionally includes pricing under various assumptions, in some cases taking a Bayesian approach (e.g., Zhang et al., 2016b; Wang et al., 2013; Devanur et al., 2019; Hong et al., 2011; Zaman and Grosu, 2012; Zhang et al., 2016a; Wang et al., 2015; Kash et al., 2019; Zhang et al., 2017; Chawla et al., 2017), but lacks our nearly-optimal worst-case guarantees for the assignment problem.

Primal-dual analysis with non-standard LP. The concept of configuration LP relaxation was initially introduced to design polynomial time approximation algorithms for the offline combinatorial optimization problems. It has been applied to various problems such as the cutting stock problem (Eisemann, 1957), the facility location problem (Jain et al., 2003), the bin packing problem (Bansal et al., 2006), combinatorial auctions (Abraham et al., 2012), and the scheduling problem (Verschae and Wiese, 2014). In many of these problems, the configuration LP achieves a smaller integrality gap compared to the standard LP relaxation (see e.g., Verschae and Wiese, 2014). In online algorithm design, there are also recent works using the configuration LP. Nguyen (2020) studies the online packing/covering problem and designs competitive algorithms with respect to the configuration LP. Huang et al. (2024) study the Adwords problem and design an online algorithm whose competitive ratio against the configuration LP beats the greedy algorithm. Huang and Zhang (2024) and Huang et al. (2023) study online bipartite matching with stochastic rewards. They use a randomized primal-dual framework to prove the competitive ratio guarantees of their algorithms with respect to the configuration LP.

In addition to the configuration LP, primal-dual analysis with other non-standard LPs has been studied in the field of online matching. Goyal and Udwani (2023) introduce a path-based formulation LP and an LP-free primal-dual analysis framework for online bipartite matching with stochastic rewards. This technique has been subsequently applied to online bipartite matching with stochastic i.i.d. reusable resources (Goyal et al., 2025), and AdWords with unknown budgets (Udwani, 2025; Manshadi et al., 2025). Huang and Shu (2021) propose a natural LP for online stochastic bipartite matching and achieved a better competitive ratio than previous results, while Aouad and Ma (2023) propose a truncated LP with a similar idea for a new variant model involving correlated arrivals.

EC.2. Competitive Ratio Analysis for Real-valued Durations

In this subsection, we present the formal analysis for the competitive ratio upper bound of FLB in Theorem 3.1.

Theorem 3.1 (Competitive Ratio) *There exists a choice of parameters $(\gamma^*, \eta^*, \beta^*)$ such that the asymptotic competitive ratio of* FLB*, with inspection-frequency scalar γ^* and penalty parameters (η^*, β^*), is at most*

$$\ln(RD) + 3\ln\ln(R \vee D) + O(1).$$

Moreover, there exist instances involving a single server with $R, D \geq 1$, for which the asymptotic competitive ratio of any online algorithm (possibly fractional or randomized) against the optimal offline benchmark is at least $\ln(RD) + \Omega(1)$. As a corollary, $\mathcal{C}_{\text{LOG-LIMIT}}(\mathsf{FLB}) = \mathcal{C}^*_{\text{LOG-LIMIT}} = 1$.

Similar to the integer-duration environment studied in Section 4, in Section EC.2.1, we first characterize a sufficient condition for the capacity feasibility of FLB. We then upper bound the competitive ratio of capacity feasible FLB as a function of its parameters (γ, η, β) in Section EC.2.2. Finally, in Section EC.2.3, we show an appropriate selection of parameters that obtains the competitive ratio upper bound in Theorem 3.1.

EC.2.1. Sufficient Condition for the Capacity Feasibility

We present the following sufficient condition for the capacity feasibility of FLB.

Proposition 4.10 (Capacity Feasibility of FLB) FLB *is capacity feasible if integer-valued inspection-frequency scalar $\gamma \in \mathbb{N}$ and penalty parameters $(\eta, \beta) \in \mathbb{R}^2_+$ satisfy*

$$\ln(\beta) \geq -\ln\left(\prod\nolimits_{k\in[\lceil\gamma D\rceil]}\left(1 - \frac{R}{k(R+\gamma\eta)}\right)\right) - \ln\left(1 + \left(\gamma + \frac{R}{\eta}\right)\left(1 - \frac{(\eta\gamma + R)\ln(\beta)}{Rc_{\min}}\right) - \frac{R}{\eta}\left(1 + \frac{\eta}{R+\gamma\eta}\right)^{\gamma}\right) + \ln\left(\frac{(\gamma+1)(R+\gamma\eta)}{\gamma\eta}\right) + \ln\left(\prod\nolimits_{k\in[\gamma+1]}\left(1 - \frac{R}{k(R+\gamma\eta)}\right)\right)$$

Similar to Proposition 4.2 following from Lemma 4.4, we prove the following invariant for the real-valued case and we use the same machinery to prove Proposition 4.10.

Lemma EC.2.1 *Consider a hypothetical scenario in which* FLB *can assign a job to a server that has no available capacity (which could result in a negative projected capacity level). With integer-valued inspection-frequency scalar $\gamma \in \mathbb{N}$, the projected available capacity under* FLB *satisfies the following property: for every server $i \in [n]$, time points $t, \tau \in \mathbb{R}_+$ such that $\gamma(\tau - t) \in \mathbb{N}_+$ and index $k \in \mathbb{N}$:*

$$\alpha_{i,t\to\tau+\frac{k}{\gamma}} - \alpha_{i,t\to\tau} \leq \begin{cases} -\dfrac{\ln\left(1 - \frac{R}{R+\eta\gamma+\eta}\left(\frac{R+\eta\gamma+\eta}{R+\eta\gamma}\right)^k - \epsilon\right)}{\ln(\beta)} & k \leq \gamma \\ -\dfrac{\ln(\Lambda) + \ln\left(\rho_{\frac{R}{\eta\gamma+R}}(k)\right)}{\ln(\beta)} & k \geq \gamma + 1 \end{cases}$$

where $\ln(\Lambda) = \ln\left(1 + \left(\gamma + \frac{R}{\eta}\right)(1-\epsilon) - \frac{R}{\eta}\left(\frac{R+\eta\gamma+\eta}{R+\eta\gamma}\right)^{\gamma}\right) - \ln\left(\frac{\gamma+1}{\frac{\eta\gamma}{\eta\gamma+R}}\rho_{\frac{R}{\eta\gamma+R}}(\gamma+1)\right)$, *and* $\epsilon = \frac{(\eta\gamma+R)\ln(\beta)}{Rc_i}$, *and* $\rho_{\frac{R}{\eta\gamma+R}}(k) = \prod_{\ell\in[k]}\left(1 - \frac{R}{\ell(R+\gamma\eta)}\right)$.

Proof of lemma EC.2.1. This holds initially at time 0 as we have $\alpha_{i,0\to t}=1$ for all t. Assume this holds true until arrival time of customer j at t_j. We will prove this inequality holds true after the possible new assignment. The only interesting and non-trivial case is when the algorithm decides to assign job j to server i and $t_j \le t < t_j + d_{ij} \le t + \frac{k}{\gamma}$ which means $Rd_{ij} - \sum_{k'\in[0:\lceil\gamma d_{ij}\rceil-1]} \Psi\left(\alpha_{i,t_j\to t_j+\frac{k'}{\gamma}}\right) > 0$ or:

$$\begin{aligned}
\frac{Rd_{ij}}{\eta} &> \frac{1}{\eta}\sum\nolimits_{k'\in[0:\lceil\gamma d_{ij}\rceil-1]} \Psi\left(\alpha_{i,t_j\to t_j+\frac{k'}{\gamma}}\right) \\
&= \frac{1}{\eta}\sum\nolimits_{k'\in[0:\gamma(\tau-t_j)]} \Psi\left(\alpha_{i,t_j\to t_j+\frac{k'}{\gamma}}\right) + \frac{1}{\eta}\sum\nolimits_{k'\in[\lceil\gamma d_{ij}\rceil+\gamma(t_j-\tau)-1]} \Psi\left(\alpha_{i,t_j\to\tau+\frac{k'}{\gamma}}\right) \\
&\ge (1+\gamma(t-t_j))\left(\beta^{(1-\alpha_{i,t_j\to\tau})}-1\right) + \sum\nolimits_{k'\in[\lceil\gamma d_{ij}\rceil+\gamma(t_j-t)-1]} \left(\beta^{(1-\alpha_{i,t_j\to\tau+\frac{k'}{\gamma}})}-1\right)
\end{aligned}$$

Case (i): $t_j + d_{ij} \le t+1$:

$$\begin{aligned}
&\overset{(a)}{\ge} -\gamma d_{ij} + (1+\gamma(t-t_j))\,\beta^{(1-\alpha_{i,t_j\to\tau})} + \sum\nolimits_{k'\in[\lceil\gamma d_{ij}\rceil+\gamma(t_j-t)-1]} \beta^{(1-\alpha_{i,t_j\to\tau}+\frac{\ln\left(1-\frac{R}{R+\eta\gamma+\eta}\left(\frac{R+\eta\gamma+\eta}{R+\eta\gamma}\right)^{k'}-\epsilon\right)}{\ln(\beta)})} \\
&= -\gamma d_{ij} + \beta^{(1-\alpha_{i,t_j\to\tau})}\left(1+\gamma(t-t_j) + \sum\nolimits_{k'\in[\lceil\gamma d_{ij}\rceil+\gamma(t_j-t)-1]}\left(1-\frac{R}{R+\eta\gamma+\eta}\left(\frac{R+\eta\gamma+\eta}{R+\eta\gamma}\right)^{k'}-\epsilon\right)\right) \\
&\ge -\gamma d_{ij} + \beta^{(1-\alpha_{i,t_j\to\tau})}\left(\gamma d_{ij}(1-\epsilon) - \frac{R}{R+\eta\gamma+\eta}\sum\nolimits_{k'\in[\lceil\gamma d_{ij}\rceil+\gamma(t_j-t)-1]}\left(\frac{R+\eta\gamma+\eta}{R+\eta\gamma}\right)^{k'}\right) \\
&\ge -\gamma d_{ij} + \beta^{(1-\alpha_{i,t_j\to\tau})}\left(\gamma d_{ij}(1-\epsilon) - \frac{Rd_{ij}}{R+\eta\gamma+\eta}\sum\nolimits_{k'\in[k-1]}\left(\frac{R+\eta\gamma+\eta}{R+\eta\gamma}\right)^{k'}\right) \\
&= -\gamma d_{ij} + \beta^{(1-\alpha_{i,t_j\to\tau})}\left(\gamma d_{ij}(1-\epsilon) - \frac{Rd_{ij}}{R+\eta\gamma+\eta}\,\frac{\left(\frac{R+\eta\gamma+\eta}{R+\eta\gamma}\right)^{k} - \frac{R+\eta\gamma+\eta}{R+\eta\gamma}}{\frac{\eta}{R+\eta\gamma}}\right) \\
&= -\gamma d_{ij} + \beta^{(1-\alpha_{i,t_j\to\tau})}\left(\left(\gamma+\frac{R}{\eta}\right)d_{ij} - \gamma d_{ij}\epsilon - \frac{Rd_{ij}}{R+\eta\gamma+\eta}\,\frac{\left(\frac{R+\eta\gamma+\eta}{R+\eta\gamma}\right)^{k}}{\frac{\eta}{R+\eta\gamma}}\right)
\end{aligned}$$

Hence:

$$\begin{aligned}
&1 \ge \beta^{(1-\alpha_{i,t_j\to\tau})}\left(1 - \frac{\eta\gamma}{\eta\gamma+R}\epsilon - \frac{R}{R+\eta\gamma+\eta}\left(\frac{R+\eta\gamma+\eta}{R+\eta\gamma}\right)^{k}\right) \\
\iff &-\frac{\ln\left(1-\frac{\eta\gamma}{\eta\gamma+R}\epsilon - \frac{R}{R+\eta\gamma+\eta}\left(\frac{R+\eta\gamma+\eta}{R+\eta\gamma}\right)^{k}\right)}{\ln(\beta)} \ge 1-\alpha_{i,t_j\to\tau}
\end{aligned}$$

On the other hand:

$$\begin{aligned}
-\frac{\ln\left(1-\epsilon-\frac{R}{R+\eta\gamma+\eta}\left(\frac{R+\eta\gamma+\eta}{R+\eta\gamma}\right)^{k}\right)}{\ln(\beta)} &\ge -\frac{\ln\left(1-\frac{\eta\gamma}{\eta\gamma+R}\epsilon-\frac{R}{R+\eta\gamma+\eta}\left(\frac{R+\eta\gamma+\eta}{R+\eta\gamma}\right)^{k}\right)-\frac{R}{\eta\gamma+R}\epsilon}{\ln(\beta)} \\
&\ge 1-\alpha_{i,t_j\to\tau}+\frac{1}{c_i}
\end{aligned}$$

Case (ii): $t_j + d_{ij} > t+1$:

$$\overset{(a)}{\geq} -\gamma d_{ij} + (1+\gamma(t-t_j))\,\beta^{(1-\alpha_{i,t_j\to\tau})} + \sum\nolimits_{k'\in[\gamma]} \beta^{(1-\alpha_{i,t_j\to\tau}+\frac{\ln\left(1-\frac{R}{R+\eta\gamma+\eta}\left(\frac{R+\eta\gamma+\eta}{R+\eta\gamma}\right)^{k'}-\epsilon\right)}{\ln(\beta)})}$$
$$+\sum\nolimits_{k'\in[\gamma+1:\lceil\gamma d_{ij}\rceil+\gamma(t_j-t)-1]} \beta^{(1-\alpha_{i,t_j\to\tau}+\frac{\ln(\Lambda)+\ln\left(\rho_{\frac{R}{\eta\gamma+R}}(k')\right)}{\ln(\beta)})}$$

$$\begin{aligned}
&\geq -\gamma d_{ij} + \beta^{(1-\alpha_{i,t_j\to\tau})}\left(1+\gamma(1-\epsilon) - \frac{R}{R+\eta\gamma+\eta}\left(\left(\frac{R+\eta\gamma+\eta}{R+\eta\gamma}\right)^{\gamma+1} - \frac{R+\eta\gamma+\eta}{R+\eta\gamma}\right)\frac{R+\eta\gamma}{\eta}\right)\\
&\quad + \beta^{(1-\alpha_{i,t_j\to\tau})}\left(\Lambda\sum\nolimits_{k'\in[\gamma+1:\lceil\gamma d_{ij}\rceil-1]}\rho_{\frac{R}{\eta\gamma+R}}(k')\right)\\
&= -\gamma d_{ij} + \beta^{(1-\alpha_{i,t_j\to\tau})}\left(1+\gamma(1-\epsilon)+\frac{R}{\eta}-\frac{R}{\eta}\left(\frac{R+\eta\gamma+\eta}{R+\eta\gamma}\right)^{\gamma} + \Lambda\sum\nolimits_{k'\in[\gamma+1:\lceil\gamma d_{ij}\rceil-1]}\rho_{\frac{R}{\eta\gamma+R}}(k')\right)\\
&\overset{(b)}{=} -\gamma d_{ij} + \beta^{(1-\alpha_{i,t_j\to\tau})}\left(1+\gamma(1-\epsilon)+\frac{R}{\eta}-\frac{R}{\eta}\left(\frac{R+\eta\gamma+\eta}{R+\eta\gamma}\right)^{\gamma}\right)\\
&\quad + \beta^{(1-\alpha_{i,t_j\to\tau})}\Lambda\left(\frac{\lceil\gamma d_{ij}\rceil}{\frac{\eta\gamma}{\eta\gamma+R}}\rho_{\frac{R}{\eta\gamma+R}}(\lceil\gamma d_{ij}\rceil) - \frac{\gamma+1}{\frac{\eta\gamma}{\eta\gamma+R}}\rho_{\frac{R}{\eta\gamma+R}}(\gamma+1)\right)\\
&\geq -\gamma d_{ij} + \beta^{(1-\alpha_{i,t_j\to\tau})}\left(\Lambda\left(\frac{d_{ij}}{\frac{\eta}{\eta\gamma+R}}\rho_{\frac{R}{\eta\gamma+R}}(k)\right)+\frac{R}{\eta}\epsilon\right)
\end{aligned}$$

which implies:

$$1 \geq \beta^{(1-\alpha_{i,t_j\to\tau})}\left(\Lambda\rho_{\frac{R}{\eta\gamma+R}}(\lceil\gamma d_{ij}\rceil)+\frac{R}{\eta\gamma+R}\epsilon\right) \iff -\frac{\ln\left(\Lambda\rho_{\frac{R}{\eta\gamma+R}}(\lceil\gamma d_{ij}\rceil)+\frac{R}{\eta\gamma+R}\epsilon\right)}{\ln(\beta)} \geq 1-\alpha_{i,t_j\to\tau}$$

Finally we get:

$$-\frac{\ln(\Lambda)+\ln\left(\rho_{\frac{R}{\eta\gamma+R}}(\lceil\gamma d_{ij}\rceil)\right)}{\ln(\beta)} \geq -\frac{\ln\left(\Lambda\rho_{\frac{R}{\eta\gamma+R}}(\lceil\gamma d_{ij}\rceil)+\frac{R}{\eta\gamma+R}\epsilon\right)-\frac{R}{\eta\gamma+R}\epsilon}{\ln(\beta)} \geq 1-\alpha_{i,t_j\to\tau}+\frac{1}{c_i}$$

Inequalities (a) holds due to the induction hypothesis and equality (b) holds due to Lemma 4.5 □

EC.2.2. Primal-Dual Analysis of the Competitive Ratio

We now present the following upper bound of capacity feasible FLB's competitive ratio.

Proposition 4.11 (Competitive Ratio of Capacity Feasible FLB) *For every $\gamma \geq 2$, $\eta > 0$ and $\beta \geq e$, the competitive ratio of a capacity feasible FLB is at most*

$$\frac{\gamma}{\gamma-1}\cdot\ln(\beta)\cdot\left(1+\gamma\eta\left(1+\beta\left(\beta^{\frac{1}{c_{\min}}}-1\right)\right)\right)$$

Proof of Proposition 4.11. In this proof, we upper bound the competitive ratio of capacity feasible FLB with inspection-frequency scalar $\gamma \in (0,1]$ and penalty parameters $\eta > 0, \beta \geq e$. To simplify the notation, we rewrite the inspection time subset $\mathcal{T}_{ij}(\gamma)$ as $\mathcal{T}_{ij}$.

Recall the dual program of the configuration LP $\mathcal{P}_{\text{OPT}}$,

$$\begin{aligned}\min_{\boldsymbol{\lambda},\boldsymbol{\theta}\geq\mathbf{0}} \sum_{j\in[m]}\lambda(j)+\sum_{i\in[n]}c_i\theta(i) \qquad &\text{s.t.}\\ \sum_{j\in S}\lambda(j)+\theta(i)\geq\sum_{j\in S}r_{ij}d_{ij} \quad & i\in[n],S\in\mathcal{S}_i\end{aligned}$$

We construct a dual solution based on the assignment decision made in FLB as follows. First, set $\lambda(j)\leftarrow 0$ and $\theta(i)\leftarrow 0$ for every job $j\in[m]$ and server $i\in[n]$. Now consider every assignment of FLB. For each job $j\in[m]$, if FLB assigns server i to job j, update the dual variables as follows:

$$\begin{aligned}\lambda(j)&\leftarrow r_{ij}d_{ij}-\sum_{\tau\in\mathcal{T}_{ij}}\Psi\left(\alpha_{i,t\to\tau}\right)\\ \theta(i)&\leftarrow\theta(i)+\sum_{\tau\in\mathcal{T}_{ij}}\left(\Psi\left(\alpha_{i,t_j\to\tau}-\frac{1}{c_i}\right)-\Psi\left(\alpha_{i,t\to\tau}\right)\right)\end{aligned}$$

Note the dual solution construction is well-defined. In particular, since FLB is capacity feasible, when FLB assigns server i to job j, there exists at least one available unit of server i. Namely, $\alpha_{i,t\to t}\geq\frac{1}{c_i}$. Moreover, for every $\tau\in\mathcal{T}_{ij}$, we have $\alpha_{i,t_j\to\tau}\geq\alpha_{i,t\to t}\geq\frac{1}{c_i}$, since $\alpha_{i,t_j\to\tau}$ is increasing in τ and $\tau\geq t_j$.

The rest of the proof is done in two steps.

[*Step i*] *Comparing objective values in primal and dual.* Here we show that the total reward of FLB is a $\ln(\beta)(1+\gamma\eta(1+\beta)(\beta^{\frac{1}{c_{\min}}}-1))$-approximation of the objective value of the constructed dual solution. To show this, we analyze the increment of the reward in FLB and the increment of the objective value of the dual solution due to every assignment decision in FLB separately.

Suppose FLB assigns server i to job j. The increment of the total reward in FLB is

$$\Delta(\text{Primal})=r_{ij}d_{ij}$$

and the increment of the objective value of the constructed dual solution can be upper bounded as follows,

$$\begin{aligned}\Delta(\text{Dual})&=r_{ij}d_{ij}-\sum_{\tau\in\mathcal{T}_{ij}}\Psi\left(\alpha_{i,t_j\to\tau}\right)+\sum_{\tau\in\mathcal{T}_{ij}}c_i\left(\Psi\left(\alpha_{i,t_j\to\tau}-\frac{1}{c_i}\right)-\Psi\left(\alpha_{i,t_j\to\tau}\right)\right)\\
&\overset{(a)}{\leq}\ln(\beta)\left(r_{ij}d_{ij}-\sum_{\tau\in\mathcal{T}_{ij}}\Psi\left(\alpha_{i,t_j\to\tau}\right)\right)-\sum_{\tau\in\mathcal{T}_{ij}}c_i\left(\Psi\left(\alpha_{i,t_j\to\tau}\right)-\Psi\left(\alpha_{i,t_j\to\tau}-\frac{1}{c_i}\right)\right)\\
&=r_{ij}d_{ij}\ln(\beta)-\sum_{\tau\in\mathcal{T}_{ij}}\left(\ln(\beta)\Psi\left(\alpha_{i,t_j\to\tau}\right)+c_i\left(\Psi\left(\alpha_{i,t_j\to\tau}\right)-\Psi\left(\alpha_{i,t_j\to\tau}-\frac{1}{c_i}\right)\right)\right)\\
&\overset{(b)}{\leq}r_{ij}d_{ij}\ln(\beta)-\sum_{\tau\in\mathcal{T}_{ij}}\left(\ln(\beta)\Psi\left(\alpha_{i,t_j\to\tau}\right)+\Psi'\left(\alpha_{i,t_j\to\tau}-\frac{1}{c_i}\right)\right)\\
&\overset{(c)}{\leq}r_{ij}d_{ij}\ln(\beta)+\sum_{\tau\in\mathcal{T}_{ij}}\eta\ln(\beta)\left(1+\beta\left(\beta^{\frac{1}{c_{\min}}}-1\right)\right)\\
&=r_{ij}d_{ij}\ln(\beta)+|\mathcal{T}_{ij}|\eta\ln(\beta)\left(1+\beta\left(\beta^{\frac{1}{c_{\min}}}-1\right)\right)\\
&\overset{(d)}{\leq}r_{ij}d_{ij}\ln(\beta)+r_{ij}\gamma d_{ij}\eta\ln(\beta)\left(1+\beta\left(\beta^{\frac{1}{c_{\min}}}-1\right)\right)\\
&=\Delta(\text{Primal})\cdot\ln(\beta)\left(1+\gamma\eta\left(1+\beta\left(\beta^{\frac{1}{c_{\min}}}-1\right)\right)\right)\end{aligned}$$

In the above derivation, inequality (a) holds since $\ln(\beta) \geq 1$ for every $\beta \geq e$ and server i's reduced reward $r_{ij}d_{ij} - \sum_{\tau\in\mathcal{T}_{ij}} \Psi\left(\alpha_{i,t_j\to\tau}\right) > 0$ implied by the assumption that FLB assigns server i to job j. Inequality (b) holds since penalty function Ψ is convex and thus $c_i(\Psi(\alpha_{i,t_j\to\tau}) - \Psi(\alpha_{i,t_j\to\tau} - \frac{1}{c_i})) \geq c_i(\alpha_{i,t_j\to\tau} - (\alpha_{i,t_j\to\tau} - \frac{1}{c_i}))\Psi'(\alpha_{i,t_j\to\tau} - \frac{1}{c_i})) = \Psi'(\alpha_{i,t_j\to\tau} - \frac{1}{c_i}))$. Inequality (c) holds by algebra as follows,

$$
\begin{aligned}
&\ln(\beta)\,\Psi\left(\alpha_{i,t_j\to\tau}\right) + \Psi'\left(\alpha_{i,t_j\to\tau} - \frac{1}{c_i}\right) \\
&= \left(\ln(\beta)\,\Psi\left(\alpha_{i,t_j\to\tau}\right) + \Psi'\left(\alpha_{i,t_j\to\tau}\right)\right) + \left(\Psi'\left(\alpha_{i,t_j\to\tau} - \frac{1}{c_i}\right) - \Psi'\left(\alpha_{i,t_j\to\tau}\right)\right) \\
&= -\eta\ln(\beta) - \eta\ln(\beta)\,\beta^{\left(1-\frac{\alpha_{i,t_j\to\tau}}{c_i}\right)}\left((\beta)^{\frac{1}{c_i}} - 1\right) \\
&\geq -\eta\ln(\beta)\left(1 + \beta\left(\beta^{\frac{1}{c_{\min}}} - 1\right)\right)
\end{aligned}
$$

Finally, inequality (d) holds since term $|\mathcal{T}_{ij}| = \lfloor \gamma d_{ij} \rfloor \leq \gamma d_{ij}$ and $r_{ij} \geq 1$

[*Step ii*] *Check the approximate feasibility of dual.* Now we show that the constructed dual solution is $(\frac{\gamma}{\gamma-1})$-approximately satisfied, i.e.,

$$\frac{\gamma}{\gamma-1}\left(\sum_{j\in S}\lambda(j) + \theta(i)\right) \geq \sum_{j\in S} r_{ij}d_{ij}$$

Fix an arbitrary server $i \in [n]$ and configuration $S \in S_i$. Note the construction of FLB and the dual solution construction guarantee that $\lambda(j)$ is weakly larger than the reduced reward of server i for every time point $j \in S$ (i.e., $\lambda(j) \geq r_{ij}d_{ij} - \sum_{\tau\in\mathcal{T}_{ij}} \Psi\left(\alpha_{i,t_j\to\tau}\right)$); and $\lambda(j) \geq 0$. Thus, the dual constraint associated with primal variable $x(i,S)$ is $(\frac{\gamma}{\gamma-1})$-approximately satisfied if the following inequality holds.

$$\frac{\gamma}{\gamma-1}\theta(i) - \sum_{j\in S}\sum_{\tau\in\mathcal{T}_{ij}} \Psi\left(\alpha_{i,t_j\to\tau}\right) \geq 0$$

We prove this inequality with a *charging argument*. At a high level, our goal is to identify values $\{\theta_{(j,\tau)}(i)\}_{j\in S,\tau\in\mathcal{T}_{ij}}$ such that

$$\forall j \in S, \tau \in \mathcal{T}_{ij}:\ \theta_{(j,\tau)}(i) - \Psi(\alpha_{i,t_j\to\tau}) \geq 0 \tag{EC.1}$$

$$\sum_{j\in S}\sum_{\tau\in\mathcal{T}_{ij}} \theta_{(j,\tau)}(i) \leq \frac{\gamma}{\gamma-1}\theta(i) \tag{EC.2}$$

The actual charging argument is as follows.

With slight abuse of notation, we introduce an auxiliary notation $\mathcal{T}_{iS} \triangleq \{(j,\tau) : j \in S, \tau \in \mathcal{T}_{ij}\}$ for every server $i \in [n]$ and configuration $S \in \mathcal{S}_i$. Recall that for every configuration $S \in \mathcal{S}_i$, the durations of two different jobs $j_1, j_2 \in S$ do not overlap and thus $\mathcal{T}_{ij_1} \cap \mathcal{T}_{ij_2} = \emptyset$. Therefore, for every two $(j_1,\tau_1), (j_2,\tau_2) \in \mathcal{T}_{iS}$, if $\tau_1 \neq \tau_2$, then $j_1 \neq j_2$.

We further partition set $\mathcal{T}_{iS}$ into $\mathcal{T}_{iS}^{(a)} \sqcup \mathcal{T}_{iS}^{(b)}$ as follows,

$$\mathcal{T}_{iS}^{(a)} \triangleq \mathcal{T}_{iS} \backslash \mathcal{T}_{iS}^{(a)}$$

$$\mathcal{T}_{iS}^{(b)} \triangleq \{(j,\tau)\in\mathcal{T}_{iS} : \tau = \max\mathcal{T}_{ij} \text{ and } |\mathcal{T}_{ij}|\geq 2\}$$

In words, set $\mathcal{T}_{iS}^{(b)}$ contains all pairs of (j,τ) if τ is the last inspection time from $\mathcal{T}_{ij}$ and $|\mathcal{T}_{ij}|\geq 2$. As a sanity check, for every $j\in S$, there exists at most one τ such that $(j,\tau)\in\mathcal{T}^{(b)}(i,t)$.

We construct a one-to-many *correspondence* σ mapping from $\mathcal{T}_{iS}^{(a)}$ to $[m]\times[T]$ that satisfies the following three properties.

- SEPARATION: for every two different $(j_1,\tau_1),(j_2,\tau_2)\in\mathcal{T}_{iS}$, $\sigma(j_1,\tau_1)\cap\sigma(j_2,\tau_2)=\emptyset$.
- FEASIBILITY-I: for every $(j,\tau)\in\mathcal{T}_{iS}$, $|\sigma(j,\tau)| = c_i - c_i\alpha_{i,t_j\to\tau}$. Moreover, for every two different $(j_1,\tau_1),(j_2,\tau_2)\in\sigma(j,\tau)$, $j_1\neq j_2$.
- FEASIBILITY-II: for every $(j,\tau)\in\mathcal{T}_{iS}$, and every $(j',\tau')\in\sigma(j,\tau)$, $t'<t$, $\tau'\leq\tau$, $\tau'\in\mathcal{T}_{ij'}$, and a unit of server i is assigned to job j' with duration $d_{ij'}>\tau-t_{j'}$.

Before showing the existence of the desired correspondence σ, we first illustrate how to construct values $\{\theta_{(j,\tau)}(i)\}_{(j,\tau)\in\mathcal{T}_{iS}}$ given correspondence σ. We first construct $\theta_{(j,\tau)}(i)$ for every $(j,\tau)\in\mathcal{T}_{iS}^{(a)}$. Specifically, for every $(j,\tau)\in\mathcal{T}_{iS}^{(a)}$, define

$$\theta_{(j,\tau)}(i) \triangleq \sum\nolimits_{(j',\tau')\in\sigma(j,\tau)} \left(\Psi\left(\alpha_{i,t_{j'}\to\tau'}-\frac{1}{c_i}\right)-\Psi\left(\alpha_{i,t_{j'}\to\tau'}\right)\right)$$

Note that property SEPARATION and property FEASIBILITY-II of correspondence σ guarantee that

$$\sum\nolimits_{(j,\tau)\in\mathcal{T}_{iS}^{(a)}} \theta_{(j,\tau)}(i)\leq\theta(i) \tag{EC.3}$$

To argue condition (EC.1) for every $(j,\tau)\in\mathcal{T}_{iS}^{(a)}$, note that

$$\begin{aligned}
\theta_{(j,\tau)}(i)-\Psi(\alpha_{i,t_j\to\tau}) &= \sum\nolimits_{(j',\tau')\in\sigma(j,\tau)} \left(\Psi\left(\alpha_{i,t_{j'}\to\tau'}-\frac{1}{c_i}\right)-\Psi\left(\alpha_{i,t_{j'}\to\tau'}\right)\right)-\Psi(\alpha_{i,t_j\to\tau})\\
&\overset{(a)}{\geq} \sum\nolimits_{(j',\tau')\in\sigma(j,\tau)} \left(\Psi\left(\alpha_{i,t_{j'}\to\tau}-\frac{1}{c_i}\right)-\Psi\left(\alpha_{i,t_{j'}\to\tau}\right)\right)-\Psi(\alpha_{i,t_j\to\tau})\\
&\overset{(b)}{\geq} \sum\nolimits_{\ell\in[c_i-c_i\alpha_{i,t_j\to\tau}]} \left(\Psi\left(\frac{c_i-\ell}{c_i}\right)-\Psi\left(\frac{c_i-\ell+1}{c_i}\right)\right)-\Psi\left(\alpha_{i,t_j\to\tau}\right)\\
&\overset{(c)}{=} -\Psi(1)\overset{(d)}{=}0
\end{aligned}$$

In the above derivation, inequality (a) holds since projected capacity level $\alpha_{i,t_{j'}\to\tau'}$ is weakly increasing in τ', and property FEASIBILITY-II of correspondence σ ensures $\tau'\leq\tau$, and penalty function Ψ is convex. For inequality (b), note property FEASIBILITY-I and property FEASIBILITY-II of correspondence σ guarantee the fact that there exists $s\triangleq c_i-c_i\alpha_{i,t_j\to\tau}$ jobs $j_1<j_2<\cdots<j_s<j$, each of which is assigned a different unit of server i and holds this unit at time point τ. This fact further implies that for each $\ell\in[s]$, $\alpha_{i,t_{j_\ell}\to\tau}\leq\frac{c_i-\ell+1}{c_i}$. Invoking the convexity of penalty function Ψ completes the argument for inequality (b). Finally, inequality (c) holds by algebra and inequality (d) holds since $\Psi(1)=0$.

Next we construct $\theta_{(j,\tau)}(i)$ for every $(j,\tau)\in\mathcal{T}_{iS}^{(b)}$ with $\{\theta_{(j,\tau)}(i)\}_{(j,\tau)\in\mathcal{T}_{iS}^{(a)}}$ constructed above. Specifically, for every $(j,\tau)\in\mathcal{T}_{iS}^{(b)}$, define

$$\theta_{(j,\tau)}(i)\triangleq\sum\nolimits_{\tau'\in\mathcal{T}_{ij}\backslash\{\tau\}}\frac{\theta_{(j,\tau')}(i)}{|\mathcal{T}_{ij}\backslash\{\tau\}|}$$

Note the above construction is well-defined. In particular, $|\mathcal{T}_{ij}\backslash\{\tau\}|\geq 1$ due to the definition of set $\mathcal{T}_{iS}^{(b)}$.

To check the constructed values $\{\theta_{(j,\tau)}(i)\}_{(j,\tau)\in\mathcal{T}}$ satisfies condition (EC.2), note that

$$\begin{aligned}\sum\nolimits_{(j,\tau)\in\mathcal{T}_{iS}}\theta_{(j,\tau)}(i)&=\sum\nolimits_{(j,\tau)\in\mathcal{T}_{iS}^{(a)}}\theta_{(j,\tau)}(i)+\sum\nolimits_{(j,\tau)\in\mathcal{T}_{iS}^{(b)}}\theta_{(j,\tau)}(i)\\&=\sum\nolimits_{(j,\tau)\in\mathcal{T}_{iS}^{(a)}}\theta_{(j,\tau)}(i)+\sum\nolimits_{(j,\tau)\in\mathcal{T}_{iS}^{(b)}}\sum\nolimits_{\tau'\in\mathcal{T}_{ij}\backslash\{\tau\}}\frac{\theta_{(j,\tau')}(i)}{|\mathcal{T}_{ij}\backslash\{\tau\}|}\\&\overset{(a)}{\leq}\sum\nolimits_{(j,\tau)\in\mathcal{T}_{iS}^{(a)}}\theta_{(j,\tau)}(i)+\frac{1}{\gamma-1}\sum\nolimits_{(j,\tau)\in\mathcal{T}_{iS}^{(b)}}\sum\nolimits_{\tau'\in\mathcal{T}_{ij}\backslash\{\tau\}}\theta_{(j,\tau')}(i)\\&\overset{(b)}{\leq}\frac{\gamma}{\gamma-1}\sum\nolimits_{(j,\tau)\in\mathcal{T}_{iS}^{(a)}}\theta_{(j,\tau)}(i)\\&\overset{(c)}{\leq}\frac{\gamma}{\gamma-1}\theta(i)\end{aligned}$$

where inequality (a) holds since $|\mathcal{T}_{ij}\backslash\{\tau\}|\geq\gamma-1$ due to the definition of inspection time subset $\mathcal{T}_{ij}$ parameterized by inspection-frequency scalar γ; inequality (b) holds since for every $j\in S$ there exists at most one τ such that $(j,\tau)\in\mathcal{T}^{(b)}(i,t)$; and inequality (c) holds due to inequality (EC.3).

To argue condition (EC.1) for every $(j,\tau)\in\mathcal{T}_{iS}^{(b)}$, note that

$$\begin{aligned}\theta_{(j,\tau)}(i)-\Psi(\alpha_{i,t_j\to\tau})&\overset{(a)}{\geq}\sum\nolimits_{\tau'\in\mathcal{T}_{ij}\backslash\{\tau\}}\frac{\theta_{(j,\tau')}(i)}{|\mathcal{T}_{ij}\backslash\{\tau\}|}-\sum\nolimits_{\tau'\in\mathcal{T}_{ij}\backslash\{\tau\}}\frac{\Psi(\alpha_{i,t_j\to\tau'})}{|\mathcal{T}_{ij}\backslash\{\tau\}|}\\&=\sum\nolimits_{\tau'\in\mathcal{T}_{ij}\backslash\{\tau\}}\frac{\theta_{(j,\tau')}(i)-\Psi(\alpha_{i,t_j\to\tau'})}{|\mathcal{T}_{ij}\backslash\{\tau\}|}\overset{(b)}{\geq}0\end{aligned}$$

where inequality (a) holds due to the fact $\Psi(\alpha_{i,t_j\to\tau})\leq\Psi(\alpha_{i,t\to\tau'})$ implied by $\tau'<\tau$ and the definition of $\theta_{(j,\tau)}(i)$ for $(j,\tau)\in\mathcal{T}_{iS}^{(b)}$; and inequality (b) holds since condition (EC.1) is shown to be satisfied for $(j,\tau')\in\mathcal{T}_{iS}^{(a)}$.

Now we show the existence of correspondence σ with properties SEPARATION, FEASIBILITY-I and FEASIBILITY-II by the following explicit construction of σ. Fix an arbitrary $(j,\tau)\in\mathcal{T}_{iS}$. The definition of projected capacity level $\alpha_{i,t_j\to\tau}$ implies that there are $s\triangleq c_i-c_i\alpha_{i,t_j\to\tau}$ jobs $j_1<j_2<\cdots<j_s<j$, each of which is assigned a different unit of server i and holds this unit at time point τ. Namely, for every $\ell\in[s]$, $t_{j_\ell}+d_{ij_\ell}\geq\tau$. Moreover, there exists unique $\tau_\ell\in\mathcal{T}_{ij_\ell}$ such that $\tau_\ell\leq\tau<\tau_\ell+\frac{1}{\gamma}$. By setting $\sigma(j,\tau)\triangleq\{(j_\ell,\tau_\ell)\}_{\ell\in[m]}$, properties FEASIBILITY-I and FEASIBILITY-II are satisfied straightforwardly. To show property SEPARATION, fix arbitrary two different $(j',\tau'),(j'',\tau'')\in\mathcal{T}_{iS}^{(a)}$ such that $j'\leq j''$. It is sufficient to argue $|\tau''-\tau'|\geq\frac{1}{\gamma}$. We consider two cases separately. If $j'=j''$, for $\tau'\neq\tau''$, the construction of inspection time subset $\mathcal{T}_{ij'}$ ensures $|\tau''-\tau'|\geq\frac{1}{\gamma}$ as desired. If $j'<j''$, since both $j',j''\in S$, inspection time subsets $\mathcal{T}_{ij'}$ and $\mathcal{T}_{ij''}$ do not overlap, i.e., $t_{j'}+d_{ij'}\leq t_{j''}$, and thus $\tau'\leq t_{j'}+d_{ij'}\leq t_{j''}\leq\tau''$. Moreover, the construction of $\mathcal{T}_{iS}^{(a)}$ ensures that $\tau'\leq t_{j'}+d_{ij'}-\frac{1}{\gamma}$. Hence, $|\tau''-\tau'|=\tau''-\tau'\geq\frac{1}{\gamma}$ as desired.

Finally, since the total reward of FLB is $\ln(\beta)(1+\gamma\eta(1+\beta)(\beta^{\frac{1}{c_{\min}}}-1))$-approximation of the objective value of the constructed dual solution (Step i) and the constructed dual solution is $(\frac{\gamma}{\gamma-1})$-approximately feasible (Step ii), invoking the LP weak duality concludes the proof. □

EC.2.3. Proof of Theorem 3.1

Similar to the discussion in Section 4, we can formulate the task of optimizing parameters (γ,η,β) of FLB for its competitive ratio as the following program:

$$\begin{aligned}
&\min_{\gamma,\eta,\beta}\ \tfrac{\gamma}{\gamma-1}\cdot\ln(\beta)\cdot\left(1+\gamma\eta\left(1+\beta\left(\beta^{\frac{1}{c_{\min}}}-1\right)\right)\right) \qquad \text{s.t.}\\
&\qquad \ln(\beta)\geq -\ln\left(\prod_{k\in[\lceil\gamma D\rceil]}\left(1-\tfrac{R}{k(R+\gamma\eta)}\right)\right)\\
&\qquad\qquad -\ln\left(1+\left(\gamma+\tfrac{R}{\eta}\right)\left(1-\tfrac{(\eta\gamma+R)\ln(\beta)}{Rc_{\min}}\right)-\tfrac{R}{\eta}\left(1+\tfrac{\eta}{R+\gamma\eta}\right)^{\gamma}\right)\\
&\qquad\qquad +\ln\left(\tfrac{(\gamma+1)(R+\gamma\eta)}{\gamma\eta}\right)+\ln\left(\prod_{k\in[\gamma+1]}\left(1-\tfrac{R}{k(R+\gamma\eta)}\right)\right)\\
&\qquad \gamma\geq 2,\eta>0,\beta\geq e
\end{aligned}\qquad (\mathcal{P}_{\text{FLB-REAL}}[R,D,c_{\min}])$$

where the first constraint comes from the capacity feasibility condition in Proposition 4.10, while the objective and second constraint come from Proposition 4.11. Now we are ready to prove Theorem 3.1.

Proof of Theorem 3.1. We consider two regimes based on the magnitude of R,D separately.

Regime (i): Suppose $R\vee D=\omega(1)$. In this case, we assign $\gamma=\lceil\ln(R\vee D)\rceil\vee 2$, $\eta=\left(\frac{1}{\ln(R\vee D)}\right)^2$. We set β such that the first constraint in program $\mathcal{P}_{\text{FLB-REAL}}[R,D,\infty]$ binds. We claim that $\ln(\beta)=\ln(RD)+3\ln\ln(R\vee D)+O(1)$. To see this, consider each term on the right-hand side of the first constraint in $\mathcal{P}_{\text{FLB-REAL}}[R,D,\infty]$:

$$\begin{aligned}
-\ln\left(\prod_{k\in[\gamma D]}\left(1-\frac{R}{k(R+\gamma\eta)}\right)\right)&\overset{(a)}{=}\ln(R)+2\ln\ln(R\vee D)+\ln(D)+O(1)\\
-\ln\left(1+\gamma+\frac{R}{\eta}\left(1-\left(1+\frac{\eta}{R+\gamma\eta}\right)^{\gamma}\right)\right)&\leq 0\\
\ln\left(\frac{(\gamma+1)(R+\gamma\eta)}{\gamma\eta}\right)+\ln\left(1-\frac{R}{(R+\gamma\eta)}\right)&=\ln\ln(R\vee D)+O(1)\\
\ln\left(\prod_{k\in[2:\gamma+1]}\left(1-\frac{R}{k(R+\gamma\eta)}\right)\right)&\leq 0
\end{aligned}$$

where equality (a) follows the same reason in Theorem 4.1. Since $R\vee D=\omega(1)$, constraint $\beta\geq e$ from $\mathcal{P}_{\text{FLB-REAL}}[R,D,\infty]$ is satisfied. Therefore, the competitive ratio of FLB (with the aforementioned parameter values) under large capacity is at most

$$\frac{\gamma}{\gamma-1}\cdot\ln(\beta)\cdot(1+\gamma\eta)=\ln(RD)+3\ln\ln(R\vee D)+O(1)$$

Regime (ii): Suppose $R\vee D=O(1)$. In this case, we assign $\gamma=2$, $\eta=\frac{R}{e-1}$. Note that with $\eta=\frac{R}{e-1}$, all terms in the right-hand side of the first constraint from $\mathcal{P}_{\text{FLB-REAL}}[R,D,\infty]$ is independent of R. Moreover, it can be verified that the right-hand side is $O(1)$ since $D = O(1)$. Hence, we set β be the maximum between this right-hand side and e. To sum up, we obtain a feasible solution of $\mathcal{P}_{\text{FLB-REAL}}[R, D, \infty]$ and the the competitive ratio of FLB (with the aforementioned parameter values) is $\frac{\gamma}{\gamma-1} \cdot \ln(\beta) \cdot (1+\gamma\eta) = O(R) = O(1)$.

Finally, the special case of $R = D = 1$ is shown in Section 4.3 using program $\mathcal{P}_{\text{FLB-INT}}[R, D, \infty]$ for integer-valued durations. □

EC.2.4. Improved Competitive Ratios for Homogeneous Rewards

Similar to Proposition 4.8, we can also derive a tighter bound on the competitive ratio for instances with heterogeneous durations but homogeneous rewards (i.e., $D \geq 1$ but $R = 1$). For this result, we extend the definition of FLB with finite inspection-frequency scalar $\gamma \in [1, \infty)$ to infinite inspection-frequency scalar $\gamma = \infty$. Specifically, under $\gamma = \infty$, we generalize the definition of inspection-time subset $\mathcal{T}_{ij}(\infty) \triangleq [t_j, t_j + d_{ij})$ and (projected-utilization-based) reduced reward as

$$r_{ij}d_{ij} - \int_{\tau \in \mathcal{T}_{ij}(\infty)} \Psi\left(\alpha_{i,t_j \to \tau}\right) \cdot \mathrm{d}\tau$$

The improved competitive ratio is as follows.

Proposition EC.2.2 (Competitive Ratio of FLB for Homogeneous Rewards) *For instances with homogeneous rewards (i.e., $R = 1$), there exists a choice of parameters (η^*, β^*) such that the asymptotic competitive ratio of FLB, with inspection-frequency scalar $\gamma^* = \infty$ and penalty parameters (η^*, β^*), is at most $\ln(D) + 4$. Moreover, there exist instances involving a single server with $R = 1, D \geq 1$, for which the asymptotic competitive ratio of any online algorithm (possibly fractional or randomized) against the optimal offline benchmark is at least $\ln(D) + 2$.*

Proof of Proposition EC.2.2. (i) Let $\Psi(x) = \eta\left(e^{(1-x)} - 1\right) dx$ and $\gamma = \infty$, the capacity feasibility condition (refer to Lemma EC.2.3) simplifies to

$$1 \geq -\ln\left(1 - \frac{1}{1+\eta} e^{\frac{\eta}{1+\eta}}\right) + \frac{1}{1+\eta} \ln(D)$$

We will prove $\eta = \ln(D) + 3$ satisfies this inequality:

$$\frac{4}{\ln(D)+4} \geq -\ln\left(1 - \frac{1}{\ln(D)+4} e^{\frac{\ln(D)+3}{\ln(D)+4}}\right) \iff e^{-\frac{4}{\ln(D)+4}} \leq \left(1 - \frac{1}{\ln(D)+4} e^{\frac{\ln(D)+3}{\ln(D)+4}}\right)$$

Defining $\zeta = \frac{1}{\ln(D)+4}$ we need to prove $0 \leq e^{\zeta} - e^{-3\zeta} - e\zeta$ for $\zeta \leq \frac{1}{4}$. Taking derivative of the right hand side we will have $e^{\zeta} + 3e^{-3\zeta} - e$ which is increasing in ζ and hence less than $e^{0.25} + 3e^{-0.75} - e < 0$. This means $e^{\zeta} - e^{-3\zeta} - e\zeta$ is decreasing in ζ. Since The inequality holds for $\zeta = 0$ we are done. Finally our competitive ratio based on Proposition 4.11 will be $1 + \eta = \ln(D) + 4$.

(ii) We use proof by contradiction. Assume we have only one resource and Let $M > c$ be a large and ϵ be a very small number. The bad example consists of two phases. In phase 1 there will be M^2 arrivals from length $1 + M\epsilon$ to 1 in a decreasing fashion and then jobs in phase 2 will have an increasing length from 1 to D.

Phase 1 starts at time 0. there will be M jobs arriving with decreasing lengths for each number picked from the set $\{1 + (M - i)\epsilon | \forall i \in [M]\}$. Since $M > c$ then there will be a number $\delta = 1 + (M - 1 - i^*)\epsilon$ such that the algorithm will reject that job. Let δ be the first among all of them. At that time adversary stops phase 1 and starts the second phase by sending a continuum (M jobs of each length) of jobs at time δ with an increasing lengths. Notice that the algorithm accepted i^* jobs in phase 1 so it has $c - i^*$ left inventory. Assume the algorithm accepts job with lengths $\{y_1, ..., y_\ell\}$ where $\ell \leq c - i^*$. The moment before job y_i we have:

$$\frac{\text{Online ALG}}{\text{Optimal Offline}} = \frac{i^* + \sum_{k \in [i-1]} y_k}{c(1 + y_i)} > \frac{1}{\ln(D) + 2}$$

This implies $y_1 < \frac{i^*(\ln(D)+2)}{c} - 1$. Since $y_1 > 1$ then the last inequality also implies $i^* > \frac{c}{\ln(D)+2}$. Now we will prove by induction $y_i < \left(\frac{i^*(\ln(D)+2)}{c} - 1\right)\left(1 + \frac{(\ln(D)+2)}{c}\right)^{(i-1)}$. For $i = 1$ it directly follows from the inequality above. For $i > 1$ notice:

$$\begin{aligned}
&1 + y_i \\
&< \frac{\ln(D) + 2}{c}\left(i^* + \sum\nolimits_{k \in [i-1]} y_k\right) \\
&< \frac{\ln(D) + 2}{c}\left(i^* + \left(\frac{i^*(\ln(D) + 2)}{c} - 1\right)\frac{\left(1 + \frac{(\ln(D)+2)}{c}\right)^{(i-1)} - 1}{\frac{(\ln(D)+2)}{c}}\right) \\
&= \left(\frac{i^*(\ln(D) + 2)}{c} - 1\right)\left(1 + \frac{(\ln(D) + 2)}{c}\right)^{(i-1)} + 1
\end{aligned}$$

Also at the end of the phase 2 we have:

$$\frac{\text{Online ALG}}{\text{Optimal Offline}} = \frac{i^* + \sum_{k \in [\ell]} y_k}{c(1 + D)} > \frac{1}{\ln(D) + 2}$$

Which means:

$$\begin{aligned}
D &< \left(\frac{i^*(\ln(D) + 2)}{c} - 1\right)\left(1 + \frac{(\ln(D) + 2)}{c}\right)^{\ell} \leq \left(1 + \frac{(\ln(D) + 2)}{c}\right)^{c - \frac{c}{\ln(D)+2}} \\
&= \left(1 + \frac{(\ln(D) + 2)}{c}\right)^{\frac{c}{\ln(D)+2}\ln(D)} \leq D
\end{aligned}$$

The contradiction shows no algorithm can beat competitive ratio $\ln(D) + 2$. $\square$

Lemma EC.2.3 *Consider a hypothetical scenario in which* FLB *can assign a job to a server that has no available capacity (which could result in a negative projected capacity level). With integer-valued inspection-frequency scalar $\gamma \in \mathbb{N}$, the projected available capacity under* FLB *satisfies the following property: for every server $i \in [n]$, time points $s, t \in \mathbb{R}_+$ and real duration $d \in \mathbb{R}$:*

$$\alpha_{i,s\to t+d} - \alpha_{i,s\to t} \le \begin{cases} -\frac{\ln\left(1-\frac{R}{R+\eta}e^{\eta\frac{d}{R+\eta}}-\epsilon\right)}{\ln(\beta)} & d \le 1 \\ -\frac{\ln\left(1-\frac{R}{R+\eta}e^{\eta\frac{1}{R+\eta}}-\epsilon\right)-\frac{R}{R+\eta}\ln(d)+\ln(1-\delta)}{\ln(\beta)} & d \ge 1 \end{cases}$$

where $\epsilon = \frac{(R+\eta)\ln(\beta)}{Rc_i}$ and $\delta = \frac{D\ln(\beta)}{c_i}$.

Proof of Lemma EC.2.3. This holds trivially initially as $\alpha = 1$. Assume this holds true until arrival time of customer j at t_j. We will prove this inequality holds true after the possible new assignment. The only interesting and non-trivial case is when $s = t_j \le t < t_j + d_{ij} \le t + d$ and the algorithm decides to assign which means $Rd_{ij} + \int_{t_j}^{t_j+d_{ij}} \Psi\left(\alpha_{i,t_j\to\tau}\right) d\tau > 0$ or:

$$\begin{aligned} Rd_{ij} &> -\int_{t_j}^{t_j+d_{ij}} \Psi\left(\alpha_{i,t_j\to\tau}\right) d\tau \\ &\ge -(t-t_j)\Psi\left(\alpha_{i,t_j\to t}\right) - \int_{t}^{t_j+d_{ij}} \Psi\left(\alpha_{i,t_j\to\tau}\right) d\tau \\ &= -(t-t_j)\eta\left(1-\beta^{(1-\alpha_{i,t_j\to t})}\right) - \eta\int_{t}^{t_j+d_{ij}} \left(1-\beta^{(1-\alpha_{i,t_j\to\tau})}\right) d\tau \end{aligned}$$

Case (i): $t_j + d_{ij} \le t+1$:

$$\begin{aligned} \frac{Rd_{ij}}{\eta} &\overset{(a)}{\ge} -d_{ij} + \beta^{(1-\alpha_{i,t_j\to t})}\left(t-t_j+\int_0^{t_j+d_{ij}-t} \beta^{\frac{\ln\left(1-\frac{R}{R+\eta}e^{\eta\frac{\tau}{R+\eta}}-\epsilon\right)}{\ln(\beta)}} d\tau\right) \\ &= -d_{ij} + \beta^{(1-\alpha_{i,t_j\to t})}\left(t-t_j+\int_0^{t_j+d_{ij}-t}\left(1-\frac{R}{R+\eta}e^{\eta\frac{\tau}{R+\eta}}-\epsilon\right)d\tau\right) \\ &\ge -d_{ij} + \beta^{(1-\alpha_{i,t_j\to t})}\left(d_{ij}(1-\epsilon) - \frac{R}{\eta}\left(e^{\eta\frac{t_j+d_{ij}-t}{R+\eta}}-1\right)\right) \\ &\ge -d_{ij} + \beta^{(1-\alpha_{i,t_j\to t})}\left(d_{ij}(1-\epsilon) - \frac{R}{\eta}\left(e^{\eta\frac{d}{R+\eta}}-1\right)d_{ij}\right) \end{aligned}$$

The last inequality came from $t_j - t + d_{ij} \le d$, $d_{ij} \ge 1$ and $1 - e^{\eta\frac{d_{ij}}{R+\eta}} \ge 1 - e^{\eta\frac{d}{R+\eta}} \ge (1 - e^{\eta\frac{d}{R+\eta}})d_{ij}$. Hence:

$$1 \ge \beta^{(1-\alpha_{i,t_j\to t})}\left(1-\frac{R}{R+\eta}e^{\eta\frac{d}{R+\eta}}-\frac{\eta}{R+\eta}\epsilon\right) \iff -\frac{\ln\left(1-\frac{R}{R+\eta}e^{\eta\frac{d}{R+\eta}}-\frac{\eta}{R+\eta}\epsilon\right)}{\ln(\beta)} \ge 1-\alpha_{i,t_j\to t}$$

$$\Rightarrow -\frac{\ln\left(1-\frac{R}{R+\eta}e^{\eta\frac{d}{R+\eta}}-\epsilon\right)}{\ln(\beta)} \ge -\frac{\ln\left(1-\frac{R}{R+\eta}e^{\eta\frac{d}{R+\eta}}-\frac{\eta}{R+\eta}\epsilon\right)-\frac{R}{R+\eta}\epsilon}{\ln(\beta)}$$

$$= -\frac{\ln\left(1-\frac{R}{R+\eta}e^{\eta\frac{d}{R+\eta}}-\frac{\eta}{R+\eta}\epsilon\right)}{\ln(\beta)}+\frac{1}{c_i}\geq 1-\alpha_{i,t_j\to t}+\frac{1}{c_i}$$

Case (ii): $t_j + d_{ij} \geq t+1$:

$$R\frac{d_{ij}}{\eta}\overset{(a)}{\geq} -d_{ij}+\beta^{(1-\alpha_{i,t_j\to t})}\left(t-t_j+\int_0^1\beta^{\frac{\ln\left(1-\frac{R}{R+\eta}e^{\eta\frac{\tau}{R+\eta}}-\epsilon\right)}{\ln(\beta)}}d\tau\right.$$
$$\left.+\int_1^{t_j+d_{ij}-t}\beta^{\frac{\ln\left(1-\frac{R}{R+\eta}e^{\eta\frac{1}{R+\eta}}-\epsilon\right)-\frac{R}{R+\eta}\ln(\tau)+\ln(1-\delta)}{\ln(\beta)}}d\tau\right)$$

$$= -d_{ij}+\beta^{(1-\alpha_{i,t_j\to t})}\left(t-t_j+1-\epsilon-\frac{R}{\eta}\left(e^{\frac{\eta}{R+\eta}}-1\right)+\left(1-\frac{R}{R+\eta}e^{\frac{\eta}{R+\eta}}-\epsilon\right)(1-\delta)\int_1^{t_j+d_{ij}-t}\tau^{-\frac{R}{R+\eta}}\right)$$
$$\geq -d_{ij}+\beta^{(1-\alpha_{i,t_j\to t})}\left(\frac{R+\eta}{\eta}-\epsilon-\frac{R}{\eta}e^{\frac{\eta}{R+\eta}}+\left(1-\frac{R}{R+\eta}e^{\frac{\eta}{R+\eta}}-\epsilon\right)(1-\delta)\int_1^{d_{ij}}\tau^{-\frac{R}{R+\eta}}\right)$$
$$= -d_{ij}+\beta^{(1-\alpha_{i,t_j\to t})}\left(\frac{R+\eta}{\eta}-\epsilon-\frac{R}{\eta}e^{\frac{\eta}{R+\eta}}+\left(\frac{R+\eta}{\eta}(1-\epsilon)-\frac{R}{\eta}e^{\frac{\eta}{R+\eta}}\right)(1-\delta)\left(d_{ij}^{\frac{\eta}{R+\eta}}-1\right)\right)$$
$$\geq -d_{ij}+\beta^{(1-\alpha_{i,t_j\to t})}\left(\frac{R+\eta}{\eta}(1-\epsilon)-\frac{R}{\eta}e^{\frac{\eta}{R+\eta}}\right)\left((1-\delta)d_{ij}^{\frac{\eta}{R+\eta}}+\delta\right)$$

Hence:

$$1\geq\beta^{(1-\alpha_{i,t_j\to t})}\left(1-\epsilon-\frac{R}{R+\eta}e^{\frac{\eta}{R+\eta}}\right)\left((1-\delta)d^{-\frac{R}{R+\eta}}+\delta d^{-1}\right)$$
$$\Rightarrow -\frac{\ln\left(1-\frac{R}{R+\eta}e^{\eta\frac{1}{R+\eta}}-\epsilon\right)-\frac{R}{R+\eta}\ln(d)+\ln(1-\delta)}{\ln(\beta)}\geq 1-\alpha_{i,t_j\to\tau}$$

Since $(1-\delta)d^{\frac{-R}{R+\eta}}+\delta d^{-1}<1$ we know:

$$\ln\left((1-\delta)d^{\frac{-R}{R+\eta}}+\delta d^{-1}\right)-\ln\left((1-\delta)d^{\frac{-R}{R+\eta}}\right)\geq\delta d^{-1}\geq\frac{\ln(\beta)}{c_i}$$

Then:

$$-\frac{\ln\left(1-\frac{R}{R+\eta}e^{\eta\frac{1}{R+\eta}}-\epsilon\right)-\frac{R}{R+\eta}\ln(d)+\ln(1-\delta)}{\ln(\beta)}$$
$$\geq -\frac{\ln\left(1-\frac{R}{R+\eta}e^{\eta\frac{1}{R+\eta}}-\epsilon\right)+\ln\left((1-\delta)d^{-\frac{R}{R+\eta}}+\delta d^{-1}\right)}{\ln(\beta)}+\frac{1}{c_i}\geq 1-\alpha_{i,t_j\to\tau}+\frac{1}{c_i}$$

inequalities (a) come from the induction hypothesis. These two cases show the induction hypothesis remains true after any assignment. □

EC.3. Omitted Proofs

In this section, we provide all missing proofs in the main text.

EC.3.1. Proof of Lemma 4.5

Lemma 4.5 (Identity) *For any $d \in \mathbb{N}$ and $z \geq 0$, $\sum_{\ell\in[0:d-1]} \prod_{k\in[\ell]}(1-\frac{z}{k}) = \frac{d}{1-z}\prod_{k\in[d]}(1-\frac{z}{k})$.*

proof of Lemma 4.5. Let us prove this by induction. For $d = 1$ the equation holds trivially. For $d > 1$:

$$\sum_{\ell\in[0:d-1]} \prod_{k\in[\ell]}(1-\frac{z}{k}) = \frac{d-1}{1-z}\prod_{k\in[d-1]}(1-\frac{z}{k}) + \prod_{k\in[d-1]}(1-\frac{z}{k})$$
$$= \frac{d-z}{1-z}\prod_{k\in[d-1]}(1-\frac{z}{k}) = \frac{d}{1-z}\prod_{k\in[d]}(1-\frac{z}{k})$$

which finishes the proof of the lemma as desired. □

EC.3.2. Proof of Lemma 4.7

Lemma 4.7 *For any sequence of jobs with types $\{r_{ij}, d_{ij}\}_{(i,j)\in E}$, the total reward of the optimal offline benchmark is upper bounded by the optimal objective value of the linear program $\mathcal{P}_{\text{OPT}}$.*

proof of Lemma 4.7. For each server $i \in [n]$, unit $k \in [c_i]$, and job subset $S \in \mathcal{S}_i$, let $\mathcal{E}(i,k,S)$ represent the indicator whether unit k of server i is assigned exclusively to jobs in S in the optimal offline benchmark. Since the optimal offline benchmark is feasible, it is evident that for every job $j \in [m]$,

$$\sum_{i\in[n]}\sum_{k\in[c_i]}\sum_{S\in\mathcal{S}_i: j\in S} \mathbb{1}\{\mathcal{E}(i,k,S)\} \leq 1$$

and for every server $i \in [n]$ and its unit $k \in [c_i]$,

$$\sum_{S\in\mathcal{S}_i} \mathbb{1}\{\mathcal{E}(i,k,S)\} \leq 1$$

Therefore, $x(i,S) = \sum_{k\in[c_i]} \mathcal{E}(i,k,S)$ results in a feasible assignment for program $\mathcal{P}_{\text{OPT}}$. Moreover, the objective value under this assignment will be equal to the total reward of the optimal offline benchmark, thereby concluding the proof. □

EC.3.3. Proof of Lemma 4.9

Lemma 4.9 *Let λ_D be the solution to the equation $e^{-1} = \prod_{k\in[D]}\left(1-\frac{\lambda_D}{k}\right)$. Then $\frac{1}{\lambda_D} \leq H(D)+2$.*

proof of Lemma 4.9. Notice that:

$$1 = \sum_{k\in[D]} \ln\left(\frac{k}{k-\lambda_D}\right) = \sum_{k\in[D]} \ln\left(1+\frac{\lambda_D}{k-\lambda_D}\right) \leq \sum_{k\in[D]} \frac{\lambda_D}{k-\lambda_D} \leq \lambda_D\left(\frac{1}{1-\lambda_D}+H(D)\right)$$

solving the quadratic equation:

$$\frac{1}{\lambda_D} \leq \frac{2H(D)}{H(D)+2-\sqrt{(H(D))^2+4}} \leq H(D)+2$$

The last inequality is equivalent to $(H(D)+2)\sqrt{(H(D))^2+4} \leq (H(D))^2+2H(D)+4$ and By squaring both sides of the inequality we get $0 \leq 4(H(D))^2$ □

EC.4. Extension: Server-Dependent Heterogeneity

Our Forward-Looking BALANCE and its competitive ratio guarantees can be generalized to the extension model where different servers have different range of rewards and durations. In particular, consider the setting where each server $i \in [n]$ is associated with $(\underline{r}^{(i)}, \underline{d}^{(i)})$, and the range of reward (duration) between server i and its compatible job $j \in N^{-1}(i)$ is $[\underline{r}^{(i)}, R \cdot \underline{r}^{(i)}]$ ($[\underline{d}^{(i)}, R \cdot \underline{d}^{(i)}]$). In this setting, we generalize FLB's construction as follows: for each job $j \in [m]$ and each compatible server $i \in N(j)$, the inspection time subset $\mathcal{T}_{ij}$ given inspection-frequency scalar γ is

$$\mathcal{T}_{ij}(\gamma) \triangleq \left\{\tau \in [t_j, t_j + d_{ij}) : \exists \ell \in \mathbb{N} \text{ s.t. } \tau = t_j + \tfrac{\ell}{\gamma} \cdot \underline{d}^{(i)}\right\}$$

and the reduced reward is computed as

$$r_{ij} d_{ij} - \sum\nolimits_{\tau \in \mathcal{T}_{ij}(\gamma)} \Psi\left(\alpha_{i, t_j \to \tau}\right) \cdot \underline{r}^{(i)} \underline{d}^{(i)}$$

Finally, FLB makes the same greedy-style decision that assign job j with server i^* with the highest positive reduced reward.

For the competitive ratio results and analysis, it can be checked that Theorems 3.1 and 4.1, Propositions EC.2.2 and 4.8 as well as other technical lemmas continue to hold.

EC.5. Competitive Ratio Lower Bound of (R, D)-agnostic Algorithms

In this section, we provide hardness results of (R, D)-agnostic algorithms. Specifically, we say a (R, D)-agnostic algorithm have competitive ratio $O(f(R, D))$ for some function f if for every $R \geq 1$ and $D \geq 1$, its competitive ratio is $O(f(R, D))$ among all instances with maximum reward R and maximum duration D. We establish competitive ratio lower bounds for (R, D)-agnostic deterministic algorithms and (R, D)-agnostic random algorithms, respectively.

Proposition EC.5.1 (Deterministic (R, D)-agnostic Negative Result) *There exists no (R, D)-agnostic deterministic integral online algorithm with competitive ratio $o(RD)$, even under large capacity.*

Proof of proposition EC.5.1. Consider an example with only one server with capacity $c \in \mathbb{N}$. Assume the following instance: at time 0, c jobs arrive with with reward $r = \sqrt{\ell}$ and duration $d = \sqrt{\ell}$ for each $\ell \in [1 : L]$ in an increasing order. Here L is picked by the adversary. Since the algorithm is (R, D)-agnostic, for every ℓ, the algorithm makes the same assignment decision for jobs with $r = d = \sqrt{\ell}$ for all examples with $L \geq \ell$. Thus, let $x_\ell \in \mathbb{N}$ be the number of jobs with $r = d = \sqrt{\ell}$ accepted by the algorithm. Notice that the capacity constraint says: $\sum_{\ell=1}^{\infty} x_\ell \leq c$. This defines a game between the algorithm and the adversary. The algorithm picks x_ℓ and the adversary picks L. Since x_ℓ are non-negative integer, there is a constant $\hat{L}$ where $x_\ell = 0$ for all $\ell > \hat{L}$. Thus, if the adversary picks $L = \omega(\hat{L})$, the revenue of the algorithm is at most $O(c\hat{L}) = O(c)$, while the optimal offline benchmark is cL. Consequently, the competitive ratio is $\Omega(L) = \Omega(RD)$. □

Proposition EC.5.2 (**(R, D)-agnostic Negative Result**) *There exists no* (R, D)*-agnostic (possibly fractional or randomized) online algorithm with competitive ratio* $O(\log(RD)\log\log(RD))$*, even under large capacity.*

Proof. Consider the same example mentioned in Proposition EC.5.1. We prove the proposition statement by contradiction. Fix an arbitrary constant $\alpha > 0$. Assume there exist a series of x_ℓ such that $\sum_{r=\ell}^{\infty} x_\ell \leq c$ and for all $L \geq 2$:

$$\sum_{\ell=1}^{L} \ell \cdot x_\ell \geq \frac{\alpha}{\ln(L)\ln\ln(L)} cL$$

where the left-hand side is the revenue of the algorithm, and the right-hand side is $\frac{\alpha}{\ln(L)\ln\ln(L)}$-fraction of the revenue in the optimal offline benchmark. Multiplying above inequality by $\frac{1}{L} - \frac{1}{L+1} = \frac{1}{L(L+1)}$ and summing up for all L gives us:

$$c \geq \sum_{L=1}^{\infty} x_\ell \geq \alpha \sum_{L=2}^{\infty} \frac{1}{(L+1)\ln(L)\ln\ln(L)} c$$

which is a contradiction, since the series on right-hand side diverges. □

EC.6. Competitive Ratio Analysis for Arbitrary Initial Capacity

In this section, we illustrate the idea of choosing parameters (β, η) in FLB for arbitrary initial capacity $c_{\min}$ and its induced competitive ratio guarantee. For simplicity, we consider the integer-valued environments (similar to Section 4). Extension to real-valued durations can be done using an approach similar to Section EC.2.

Recall the optimization $\mathcal{P}_{\text{FLB-INT}}[R, D, c_{\min}]$:

$$\min_{\eta,\beta} \ln(\beta) \cdot \left(1 + \eta\left(1 + \beta\left(\beta^{\frac{1}{c_{\min}}} - 1\right)\right)\right) \qquad \text{s.t.}$$
$$\ln(\beta) \geq -\ln\left(\prod_{k\in[D]}\left(1 - \frac{R}{k(R+\eta)}\right) - \frac{(R+\eta)\ln(\beta)}{Rc_{\min}}\right)$$
$$\eta > 0, \beta \geq e.$$

First notice that the constraint can be written as:

$$-\ln\left(\frac{1}{\beta} + \frac{(R+\eta)\ln(\beta)}{Rc_{\min}}\right) \geq -\ln\left(\prod_{k\in[D]}\left(1 - \frac{R}{k(R+\eta)}\right)\right)$$

Setting $\eta = \frac{1}{\ln(R \vee D)}$ and making the constraint tight we get:

$$-\ln\left(\frac{1}{\beta} + \frac{(R\ln(R\vee D)+1)\ln(\beta)}{R\ln(R\vee D)c_{\min}}\right) = -\ln\left(\prod_{k\in[D]}\left(1 - \frac{R\ln(R\vee D)}{k(R\ln(R\vee D)+1)}\right)\right)$$

Looking at this equation as

$$-\ln\left(\frac{1}{\beta} + A\ln(\beta)\right) = B,$$

We can solve it and get

$$\frac{1}{\beta} = -AW_{-1}\left(-\frac{1}{Ae^{\frac{e^{-B}}{A}}}\right). \tag{EC.4}$$

We chose the non-principal branch because a smaller β leads to a lower objective value. For the equation to have a solution, the input to the Lambert function must be greater than $-1/e$, which is equivalent to:

$$\frac{1}{A} \geq -e^B W_{-1}\left(-e^{-B-1}\right) \qquad \text{or} \qquad \frac{1}{A} \leq -e^B W_0\left(-e^{-B-1}\right).$$

Therefore, if

$$c_{\min} \geq \frac{R\ln(R \vee D)+1}{R\ln(R \vee D)} e^B 2(B+1),$$

using the fact that $2\ln(-z) \leq W_{-1}(z) \leq \ln(-z)$ we will have:

$$\frac{1}{A} = \frac{R\ln(R \vee D)c_{\min}}{R\ln(R \vee D)+1} \geq -e^B W_{-1}\left(-e^{-B-1}\right)$$

Therefore, for a large enough $c_{\min}$ we can set β as in Eqn. (EC.4) and get:

$$\beta = \frac{1}{-AW_{-1}\left(-\frac{1}{Ae^{\frac{e^{-B}}{A}}}\right)} \leq \frac{1}{e^{-B}+A\ln(A)} = \frac{1}{e^{-B}+\frac{R\ln(R\vee D)+1}{R\ln(R\vee D)c_{\min}}\ln(\frac{R\ln(R\vee D)+1}{R\ln(R\vee D)c_{\min}})} \leq \frac{e^B}{1-2e^B\frac{\ln(c_{\min})}{c_{\min}}}.$$

Plugging this upper bound for β, $\eta = \frac{1}{\ln(R\vee D)}$, and using the inequality $\ln(1-x) \geq -\frac{x}{1-x}$, we get a competitive ratio better than:

$$\left(B-1+\frac{1}{1-2e^B\frac{\ln(c_{\min})}{c_{\min}}}\right)\left(1+\frac{1}{\ln(R\vee D)}\left(1+\frac{e^B}{1-2e^B\frac{\ln(c_{\min})}{c_{\min}}}\left(\left(\frac{e^B}{1-2e^B\frac{\ln(c_{\min})}{c_{\min}}}\right)^{\frac{1}{c_{\min}}}-1\right)\right)\right)$$

Notice that when $c_{\min} \to \infty$ the above term simplifies to

$$B\left(1+\frac{1}{\ln(R\vee D)}\right),$$

Here, both the parameter assignment and the competitive ratio converge to the ones in Section 4.3 for the large capacity regime (i.e., $c_{\min} \to \infty$).