Many-Server Queueing Systems with Heterogeneous Strategic Servers in Heavy Traffic

Burak Büke\textsuperscript{a}, Gonçalo dos Reis\textsuperscript{a,b} and Vadim Platonov\textsuperscript{a}

\textsuperscript{a} School of Mathematics, University of Edinburgh, The King’s Buildings, Edinburgh, UK
\textsuperscript{b} Centro de Matemática e Aplicações (CMA), FCT, UNL, Portugal

9\textsuperscript{th} November, 2022 (01h43)

Abstract

In most service systems, the servers are humans who desire to experience a certain level of idleness. In call centers, this manifests itself as the call avoidance behavior, where servers strategically adjust their service rate to strike a balance between the idleness they receive and effort to work harder. Moreover, being humans, each server values this trade-off differently and has different capabilities. Drawing ideas on mean-field games we develop a novel framework relying on measure-valued processes to simultaneously address strategic server behavior and inherent server heterogeneity in service systems. This framework enables us to extend the recent literature on strategic servers in four new directions by: (i) incorporating individual choices of servers, (ii) incorporating individual abilities of servers, (iii) modeling the discomfort experienced by servers due to low levels of idleness, and (iv) considering more general routing policies. Using our framework, we are able to asymptotically characterize asymmetric Nash equilibria for many-server systems with strategic servers.

In simpler cases, it has been shown that the purely quality-driven regime is asymptotically optimal. However, we show that if the discomfort increases fast enough as the idleness approaches zero, the quality-and-efficiency-driven regime and other quality driven regimes can be optimal. This is the first time this conclusion appears in the literature.

Keywords: Many-server queues, strategic servers, measure-valued processes, mean field games

2010 AMS subject classifications:
Primary: 90B22, 60F17, 91A15 Secondary: 60K25

Contents

1 Introduction 2
1.1 Our Contributions ........................................ 3
1.2 Notation ........................................ 4

2 Literature on Strategic Behavior in Queues 4

3 A Many-Server Queueing Model with Strategic and Heterogeneous Servers 5
3.1 Dynamics of the System Processes .................................... 7
3.2 The Fairness Process .................................... 8

4 Convergence of the Scaled System Length Processes 9

5 A Generalized Random Routing Policy for Purely-Quality-Driven Policies 11
1 Introduction

The design and optimization of many-server service systems has been the subject of vigorous research in the last two decades, and due to the economic importance of these systems but also the challenging problems underpinning them. With a worldwide market size of $339.4 billion in 2020 and an expected market size of $500 billion by 2027, call centers constitute a perfect example of these systems [Sta20]. Around 76% of these call centers serve institutions that generate revenues above $250 million/year and more than 13% serve institutions with an annual revenue of $25 billion [Del17]. Hence, it is no surprise that a small improvement in the service time can lead to millions of dollars in savings [GLM+10].

Despite the rapid technological developments, 82% of the US and 74% of the non-US customers want more human interaction (see [PWC18]) and the vast majority of the interactions in call centers are carried out by human agents (servers). The presence of this human factor poses interesting modeling challenges. For example, the servers should be able to experience a certain level of idleness between calls in order to avoid burnout and keep their motivation high. Failure to provide these idle times yields in servers strategically slowing down their service in order to create additional idle time – this is generally referred as call avoidance behavior [Go421]. For example, servers might choose to delay pressing the terminate call button when a call ends or is transferred, and in systems where the call is routed to the server that has been available for longest, servers can become unavailable for a short period of time to put themselves at the end of the list (see e.g., [Cal17]). As a more extreme example of call avoidance behavior, [GKM03] observe that a significant number of calls last nearly zero seconds, which indicates that some servers are strategically hanging up on the customers to create extra idle time (see also [GDWW16]).

Starting with [GDWW16], there has been an increasing research interest in the design and modelling of many-server systems with strategic servers as way to determine incentives for servers to work faster. The main stream of work concentrates on the analysis of Markovian many-server queues and aims to analyze equilibria using the stationary probabilities for such systems. Hence, the characterization of the stationary probabilities constitutes a key step in this analysis. To the best of our knowledge, the only existing result in this direction is in [Gum60], where the stationary probabilities for many-server queues with heterogeneous servers is provided under a random routing policy and where the customers are routed to one of the idle servers uniformly at random. Unfortunately, even for this special case, the equations for the stationary probabilities are extremely involved, rendering an exact analysis of the general case intractable. The recent work on strategic servers mainly
concentrates on identifying symmetric equilibria by setting all service rates equal except for the rate of a tagged server.

In addition to the strategic behavior of servers, each server is an server having inherent individual differences and preferences. Each server has a different capability to serve customers, values idle time differently and, at equilibria, one expects each server to serve with a different rate if such equilibrium exists. It is then crucial for the next generation of models to capture these behaviors and analyze the so-called asymmetric equilibria in which servers serve with heterogeneous rates and preferences.

The key step in the analysis of asymmetric equilibria is identifying how idleness is distributed among servers featuring heterogeneous service rates and such heavily depends on the routing policy used by the call center. In a recent paper, [BQ22] propose the use of a measure-valued process, referred to as the fairness process, to capture the individual differences in service rates as means to obtain access to the distribution of server idleness under a given routing policy. [BQ22] prove that in a modified version of Halfin-Whitt scaling [HW81], the sequence of fairness processes is tight and converges in probability to a constant deterministic process for some common routing policies. Then, they use the fairness process to prove diffusion limits for many-server systems with heterogeneous servers. In this work, we use the fairness process as our main technical tool to analyze many-server systems with strategic heterogeneous servers.

The literature from human resource management indicates that the perception of fairness has a significant effect on employee performance [CCW01, CCS01]. The distribution of idleness among servers is generally perceived as a measure of idleness. An interesting question regarding the distribution of idleness is whether one can design routing policies in order to distribute the idleness among servers according to servers’ service rates in a pre-specified manner. For example, it is well-known that if the fastest-server-first policy is employed, i.e., the customers are routed to the idle server with the highest service rate, only the slowest servers experience idleness [Arm05, Ata08, BQ22]. As another example, [BQ22] show that if the customers are routed using a class of policies coined totally blind policies, where the server that a customer is routed to is asymptotically independent of the service rate of said server, then the cumulative idleness servers receive is proportional to their service rate. Totally blind policies include longest-idle-server-first, where customers are routed to the server who has been available longest, and random routing as studied in [Gum00]. Several routing policies have been designed in the literature for service systems with pools of servers to ensure that each pool asymptotically receives a certain proportion of idleness [Arm05, ASS11, AW10, GW09, RS15, WA13]. However, to the best of our knowledge, the question of how to design policies to distribute idleness on an individual server level remains an open problem.

1.1 Our Contributions

Our aim is to push the frontier for the analysis of service systems with strategic servers. Our contributions can be described under three major headings.

A General Framework to Study Heterogeneous Strategic Servers. In this paper, we introduce a novel framework to study systems with strategic servers and extend the results in four major directions. First, we suggest a utility function to model the individual preferences of servers regarding the trade-off between utility of idleness and cost of working faster. Second, we address the inherent differences between the abilities of servers using random individual maximum and minimum attainable service rates. Third, we model the contribution of percentage of idleness as a concave increasing function of idleness, which better reflects the marginal benefit of idleness and allows us
to model the discomfort experienced by the servers. We also show the relationship between this discomfort and the staffing level under mild assumptions without assuming any particular routing policy. Finally, we characterize the Nash equilibria for service rate distribution under a more general random routing policy.

**Service Systems with Heterogeneity under Quality-Driven Regime.** In Section 4, we provide fluid limits for many-server queues with heterogeneous service rates under quality-driven regimes. A key question regarding fluid and diffusion approximations is whether these approximations can be used to study the stationary behavior of the systems, which require an interchange of limits argument. We prove that this is indeed the case for our approximations under both quality- and quality-and-efficiency driven regimes.

**Analysis of General Randomized Policies.** The simple random routing policy, where an incoming job is assigned to one of the idle servers uniformly at random, is well-understood. However, to the best of our knowledge, there does not exist any results in the literature for many-server systems with heterogeneous servers where incoming customers are assigned randomly to the servers with probabilities depending on the service rates. In Section 5, we asymptotically characterize the distribution of idleness among servers with heterogeneous rates under a quality-driven system. As we show in Section 4, this result can be used to analyze the stationary behavior of these many-server systems. Moreover, we also show that this result can also be used to characterize the long run percentage a server with a given service rate is idle, which requires another interchange of limits argument.

1.2 Notation

In this work, we assume that all the stochastic processes and the random variables lie in the probability space $(\Omega, \mathcal{F}, P)$. We use the shorthand notation w.p. 1 instead of with probability 1 with respect to $P$. We denote the set of real numbers and the positive integers as $\mathbb{R}$ and $\mathbb{N}$, respectively. We use $\to$ and $\Rightarrow$ to denote convergence in $\mathbb{R}$, convergence in probability and convergence in measure, respectively. To emphasize a certain system parameter $x$ is random, we use the ‘$\sim$’ notation as $\tilde{x}$. Also, we define $(x)^+ := \max\{x, 0\}$ and $(x)^- := \max\{0, -x\}$. We define $C^b_{[\alpha, \beta]}[0, \infty)$ to be the set of continuous bounded functions defined on the interval $[\alpha, \beta]$, $\iota(\cdot)$ to be the identity function and $\mathbb{I}$ to be the indicator function where $\mathbb{I}(A)$ is 1 if $A$ holds and 0 otherwise. The spaces of right-continuous functions and left-continuous functions are endowed with the Skorokhod-$J_1$. For any measure $\eta$, $\langle f, \eta \rangle := \int f(x) d\eta(x)$, and hence, if $\eta$ is a probability measure $\langle \iota, \eta \rangle$ denotes the first moment of $\eta$. With a slight abuse of notation, we use $F_a$ to denote both the law and the cumulative distribution of random variable, i.e., $F(A)$ denotes the probability of set $A$ and $F(a) = F((-\infty, a])$ for any real argument $a$. Finally, for any stochastic process $X(t)$, we use $X(\infty)$ to denote a random variable distributed with its stationary distribution. For measure-valued processes, we find that it is more appropriate to use $\eta_t$ notation instead of $\eta(t)$, and we use $\eta_\infty$ to denote the stationary version.

2 Literature on Strategic Behavior in Queues

The vast majority of the literature on queues with strategic players concentrates on the strategic behavior of customers while joining a queue. We refer the reader to the excellent monographs [HH03] and [Has20] for an extensive review of the literature on strategic customers. Motivated by the competition between two firms, two-server queues with strategic servers has also received attention in the literature (see e.g. [KKR92, GW98, CH02, CZ07]). More recently, [AG10] extend this line of research to cover multiple firms each of which are operating as $M/M/N$ queues.
The study of servers who are behaving strategically within the same organization is relatively recent and to the best of our knowledge started with the seminal work of [GDWW16]. They assume that the servers (agents) are identically sensitive to the idleness they receive and there is a non-cooperative game between servers to set their service rates in order to maximize their utility which is defined as a combination of the server’s idle time and effort to speed up the service. The system operator routes the customers to the servers with the aim of incentivizing servers using the idleness experienced – they refer to this as incentive-aware routing. In an $M/M/N$ queue setting, they show the existence of symmetric equilibrium under policies which route customers to servers according to their idle time. To solve the problem, [GDWW16] explicitly assume that the servers need to choose service rates in a way to ensure stability in the system. [ARS21] show that this assumption is not necessary using Tarski’s intersection theorem. [Gop21] studies the above results comparing queueing systems with pooled and dedicated queues under $r$-routing policies, a class of randomized policies which routes the customers to servers randomly with probabilities proportional to the $r$th power of the service rate. He provides analytical results for systems with dedicated queues, but due to their complexity he only studies the pooled queues numerically. One of the major contributions of our work is to provide analytical tools for the asymptotic study of pooled queues. In addition to incentive-aware routing, one can also provide monetary incentives based on the service speed. [ZW19] characterize a joint routing and payment policy to optimize the performance under the assumption that there is a trade-off between service speed and quality of service. [ZGW21] analyze equilibrium in loss systems with finite buffer capacity when both payment incentives and incentive-aware routing is used. An interesting recent work by [BBC19] also considers strategic servers with dedicated queues and designs a dynamic arrival rate control policy. To the best of our knowledge, this work is the first application of mean-field games to queueing systems.

3 A Many-Server Queueing Model with Strategic and Heterogeneous Servers

In this manuscript we investigate queueing systems with many servers that strategically decide on their service rates and have individual heterogeneous preferences regarding the trade-off between the idleness experienced and the effort invested to work faster. We consider a sequence of queueing systems where arrivals at the $n$th system follow a Poisson process with rate $\lambda_n$. We have the following standard assumption on arrival rates,

**Assumption 3.1.** As $n \to \infty$, $\lambda_n/n \to \bar{\lambda}$, where $0 < \bar{\lambda} < \infty$.  

If there are idle servers upon a customer’s arrival, the customer is routed to one of these servers according to a pre-specified routing mechanism. All servers in the system have the skills to serve any arriving customer, albeit with different rates. If all the servers are busy when a customer arrives, the customer waits in a queue. The customers are impatient with exponential($\gamma$) patience times, independent of other customers. If the patience time of a customer expires before the service commences, the customer abandons the system. Once the service commences, the customer only departs when the service finishes. We also assume that customers are served on a first-come-first-served basis.

For the system outlined above, the system operator decided on the routing policy and the staffing level. The routing policy determines how the idleness is distributed among servers with different rates and is prescribed by the system operator. At this point, we only assume that the routing policy is non-idling, i.e., there can only be customers in the queue if all the servers in the system are busy.
Due to the non-cooperative game between servers, there is a circular relationship between staffing level and individual service rates. The staffing level is determined by the expected service rate and the distribution of service rates is determined by the idleness experienced by each server with different rate. The latter being directly influenced by the staffing level and the routing policy. Our main goal in this work is to identify whether equilibria exist and characterize any equilibrium. As the servers differ individually in terms of their capabilities and preferences, the resulting equilibria will be asymmetric in general.

The system operator aims to set the staffing level to be the offered load, i.e., with a minimum number of staff required to stabilize the system if there were no abandonments, and a safety staffing proportional to a power $\alpha$ of the offered load, where $1/2 \leq \alpha \leq 1$, to achieve a certain quality of service. Suppose that the server $k$ in the $n$th system serves with rate $\bar{\mu}_k^n$. If the system operator knows the distribution of these service rates to be $F$ with an expected value $\bar{\mu}_F$, she sets the staffing level for the $n$th system to be

$$N^n_\alpha = \frac{\lambda^n}{\bar{\mu}_F} + \beta \left( \frac{\lambda^n}{\bar{\mu}_F} \right)^\alpha,$$

(3.1)

where $\beta > 0$ (see e.g. [GKM03]). Following the classification in [GMR02], we say that the system operates in a quality-driven (QD) regime when $1/2 < \alpha \leq 1$ and in a quality-and-efficiency-driven (QED) regime when $\alpha = 1/2$. In the literature, it is quite common to restrict the quality-driven regime to the case with $\alpha = 1$ and this case also plays a major role in our study. Hence, to differentiate, we say that the system operates in a purely quality-driven regime if the system operates with $\alpha = 1$.

On the other hand, each server in the system aims to set her service rate as to maximize her own utility, where utility is defined as a trade-off between the expected long run proportion of idleness she experiences and an effort cost she incurs for working faster. This is similar to [GDWW16] with two fundamental differences. First, we model the contribution of idleness to the utility as a concave increasing function of long run percentage idleness experienced rather than the long run percentage idleness itself. In addition to modeling the decreasing marginal return of idleness, this approach also helps us model the situations where working without any breaks is unacceptable to the servers. Second, we model the trade-off between ‘benefits of idleness’ and ‘effort cost’ to be server specific by introducing a multiplicative random coefficient to the effort cost function. The idleness each server experiences depends ultimately on the service rates of all servers in the system. Defining $I_k^n(t)$ as the idleness process of server $k$, where $I_k^n(t)$ takes the value 1 if the server $k$ in system $n$ is busy at time $t$; $I_k^n(t)$ takes the value 0 otherwise. We define the utility function of server $k$ in the $n$th system, serving with rate $\mu$, to be

$$U_k^n(\mu, F) = f(\mathbb{E}[I_k^n(\infty)|\bar{\mu}_k^n = \mu]) - \bar{a}_k^n c(\mu).$$

(3.2)

We refer to $f(\cdot)$ as the utility of idleness function and assume it to be concave and increasing. The effort cost function, $c(\cdot)$, is assumed to be convex and increasing. We also assume that both the utility of idleness function and the effort cost function are twice continuously differentiable. The coefficient $\bar{a}_k^n$ is a positive random variable with distribution $F_n(\cdot)$ and density function $f_n(\cdot)$, and determines the personal preference of server $k$ in regards to the trade-off between utility of idleness and the effort cost.

Each server $k$ in the $n$th system has an inherent minimum service rate, $\bar{\mu}_{\text{min},k}$, which the server can achieve with minimal effort and an inherent maximum achievable service rate, $\bar{\mu}_{\text{max},k}$, which the server cannot improve upon (by working faster) regardless of how much effort she puts into her work. The system operator does not have prior information about these individual minimum and maximum rates and hence, they are modeled as random variables following a common joint
distribution $F_{\min, \max}$ with marginals $F_{\min}$ and $F_{\max}$. We assume that these individual minimum and maximum service rates are uniformly bounded, i.e., there exists $\mu_{\min}$ and $\mu_{\max}$ such that

$$0 < \mu_{\min} \leq \mu_{\min, k} \leq \tilde{\mu}_k^n \leq \mu_{\max, k} \leq \mu_{\max} < \infty, \quad \text{for all } k, n \in \mathbb{N} \quad \text{w.p. } 1. \quad (3.3)$$

Without loss of generality, we assume that $\mu_{\min} = 1$ by changing the time units as necessary.

Given a distribution $F^{(1)}$ for the service rates $\{\tilde{\mu}_k^n\}$, each server $k$ in the $n$th system serves with the service rate $\tilde{\mu}_k^n$ that maximizes her utility, i.e.,

$$\tilde{\mu}_k^n \in \arg \max_{\mu \in [\mu_{\min, k}, \mu_{\max, k}]} U^n(\mu, F^{(1)}).$$

As $\tilde{\alpha}_k^n, \tilde{\mu}_{\min, k}^n$ and $\tilde{\mu}_{\max, k}^n$ are unknown to the system operator, the resulting service rate is random and the distributions of these parameters determine the distribution $F^{(1)}$ of the optimal service rates. Hence, to study a system with strategic heterogeneous servers, one needs to characterize the fixed point where $F^{(0)} = F^{(1)} = F$.

### 3.1 Dynamics of the System Processes

In this section, we provide rigorous mathematical definitions of the stochastic processes in our model. We denote the number of customers in the $n$th system at time $t$ as $X^n(t)$. For the $n$th system, the arrival process is a Poisson process with rate $\lambda^n$ and $A^n(t)$ denotes the number of arrivals by time $t$. The service time of a customer depends on the server and takes an exponential time with rate $\tilde{\mu}_k^n$ if the customer is served by server $k$. In this section, we assume that the service rates $\tilde{\mu}_k^n, 1 \leq k \leq N^n$, are i.i.d. random variables with a common known distribution $F$ having mean $\tilde{\mu}_F$ and variance $\sigma^2_F$.

In Section 6, we analyze in detail how this distribution is determined as a result of the strategic decisions of the servers. The departure process for server $k$ and the number of abandonments by time $t$ is denoted $D_k^n(t)$ and $R_k^n(t)$, respectively. We have the following balance equation for the system length process $X^n(t)$,

$$X^n(t) = X^n(0) + A^n(t) - \sum_{k=1}^{N^n} D_k^n(t) - R_k^n(t), \quad \text{for all } t \geq 0, n \in \mathbb{N}. \quad (3.4)$$

Taking $S_k^n(t)$ and $R_k^n(t)$ as independent unit rate Poisson processes for all $k$ and $n$ and considering the idleness processes $I_k^n(t)$, we can equivalently define the departure and the abandonment processes as

$$D_k^n(t) := S_k^n \left( \tilde{\mu}_k^n \int_0^t (1 - I_k^n(s)) \, ds \right) \quad \text{and} \quad R_k^n(t) := R^n \left( \gamma \int_0^t (X^n(s) - N^n)^+ \, ds \right),$$

and can write (3.4) as

$$X^n(t) = X^n(0) + A^n(t) - \sum_{k=1}^{N^n} S_k^n \left( \tilde{\mu}_k^n \int_0^t (1 - I_k^n(s)) \, ds \right) - R^n \left( \gamma \int_0^t (X^n(s) - N^n)^+ \, ds \right), \quad (3.5)$$

for all $t \geq 0, n \in \mathbb{N}$. The non-idling property of the routing policy implies

$$(X^n(t) - N^n)^- = \sum_{k=1}^{N^n} I_k^n(t), \quad \text{for all } t \geq 0, n \in \mathbb{N}.$$
For all $1/2 \leq \alpha \leq 1$, we define the following scaled processes

$$\hat{X}_\alpha^n(t) := n^{-\alpha}(X^n(t) - N^n_\alpha) \quad \text{and} \quad \hat{I}_{k,\alpha}^n(t) := n^{-\alpha}I^n_k(t).$$

We also use the shorthand notation $\hat{I}_\alpha^n(t) := \sum_{k=1}^{N^n_\alpha} \hat{I}_{k,\alpha}^n(t)$ for the total scaled idleness in the system.

Regarding the initial conditions, we make the following assumption

**Assumption 3.2.** There exists an $M_0 < \infty$ such that $|\hat{X}_\alpha^n(0)| \leq M_0$ w.p. 1 for all $n \in \mathbb{N}$. Moreover, $\hat{X}_\alpha^n(0) \Rightarrow \xi_0$ as $n \to \infty$.

The boundedness condition in Assumption 3.2 might seem more restrictive than standard assumptions in the literature. In fact, our main results can still be shown to hold when this condition is relaxed to the collection of random variables $\{\hat{X}_\alpha^n(0)\}_{n \in \mathbb{N}}$ being uniformly integrable. However, this level of generality significantly complicates the notation and statement of some of the results below, and hence we choose to restrict ourselves to boundedness assumption for exposition purposes.

### 3.2 The Fairness Process

The utility of a server is defined based on the long run proportion of time the server stays idle. This proportion is determined by the total idleness the system experiences and how this idleness is distributed among servers as a consequence of the system operator’s choice of routing policy.

As we show in Section 4, the total idleness is relatively easier to analyze once the distribution of idleness is characterized. However, the distribution of idleness is more difficult to analyze and constitutes the major challenge in identifying the utility of servers. Hence, the recent literature (see e.g., [GDWW16, ARS21]) has concentrated on servers with identical preferences and has aimed to identify a symmetric equilibrium by considering a system where all the servers, except the one under consideration, have the same rate.

To be able to address strategic servers with heterogeneous service rates, the tool we use is the fairness process introduced by [BQ22]. They define the fairness process as a probability measure-valued process showing how the idleness is distributed among servers with different service rates.

The name is motivated by the fact that the distribution of idleness is generally viewed as a measure of fairness towards servers (see e.g., [AW10, WA13]). To define the fairness process rigorously, set $\tau^n_\epsilon := \inf\{t > 0 : \int_0^t \hat{I}_\alpha^n(s)ds > \epsilon\}$ as the first time the cumulative scaled idleness experienced by the system is greater than $\epsilon \geq 0$. Then, the fairness process for the $n$th system is defined as

$$\eta^n_{\alpha,t}(A) := \begin{cases} \frac{\sum_{k=1}^{N^n_\alpha} \delta_{\mu^n_k(A)} \int_0^t \hat{I}_{k,\alpha}^n(s)ds}{\delta_0(A)} & \text{if } t > \tau^n_\epsilon, \\ \int_0^t \hat{I}_\alpha^n(s)ds & \text{if } t \leq \tau^n_\epsilon, \end{cases} \quad (3.6)$$

for all $A \in \mathcal{B}(\mathbb{R}_+)$. Unfortunately, it can easily be shown that the fairness processes do not converge in measure in any of the four Skorokhod topologies as $n \to \infty$. Hence, [BQ22] suggest using shifted versions of the fairness process, i.e., the $\epsilon$-shifted fairness processes $S_{\epsilon,\eta^n_{\alpha,t}}$,

$$S_{\epsilon,\eta^n_{\alpha,t}}(A) := \begin{cases} \frac{\sum_{k=1}^{N^n_\alpha} \delta_{\mu^n_k(A)} \int_0^t \hat{I}_{k,\alpha}^n(s)ds}{\delta_0(A)} & \text{if } t > \tau^n_\epsilon, \\ \int_0^t \hat{I}_\alpha^n(s)ds & \text{if } t \leq \tau^n_\epsilon, \end{cases} \quad (3.7)$$

for all $A \in \mathcal{B}(\mathbb{R}_+)$. The following lemma shows that the scaled idleness processes are stochastically bounded. This result is key in enabling us to study the convergence of fairness measures as well as derive fluid and diffusion limits in Section 4.
Lemma 3.3. For $1/2 \leq \alpha \leq 1$, the sequence of processes $\{(\hat{X}_n^n(t))^{+}\}_{n \in \mathbb{N}}$ and $\{\hat{I}_n^n(t)\}_{n \in \mathbb{N}}$ are stochastically bounded. Moreover, for $1/2 < \alpha \leq 1$ there exists a time $t_M > 0$ such that $\sup_{t \leq t_M} (\hat{X}_n^n(t))^{+} \xrightarrow{n \to \infty} 0$ for all $T > t_M$ as $n \to \infty$.

Using Lemma 3.3, it is straightforward to prove tightness of the continuous processes $\left\{ \int_0^t \hat{I}_n^n(s) ds \right\}_{n \in \mathbb{N}}$, which in turn implies the tightness of $\{\tau_n\}_{n \in \mathbb{N}}$ by an appeal to the continuous mapping theorem. The second part of Lemma 3.3 states that in quality-driven systems, the positive part of the scaled system length process $\{(X_\alpha^n(t))^{+}\}_{n \in \mathbb{N}}$ converges (in $n$) to 0 after time $t_M$. As we are generally concerned with the steady-state behavior of the system, we henceforth assume that $\hat{X}_n^n(0) \leq 0$ for all $n \in \mathbb{N}$. We are now ready to define the limiting fairness process.

Definition 3.4. Suppose $\tau^n_\epsilon \Rightarrow \tau_\epsilon$ for all $\epsilon > 0$ as $n \to \infty$. A measure-valued process $\{\eta_{\alpha,t}\}_{t \geq 0}$ is called the limiting fairness process if for all $\epsilon > 0$, it holds that $S_{t} \eta_{\alpha,t}^{n} \Rightarrow S_{t} \eta_{\alpha,t}$ as $n \to \infty$ in the Skorokhod-$J_1$ topology modified for left-continuous functions, and where $S_{t} \eta_{\alpha,t}$ is defined by replacing $\tau^n_\epsilon$ in (3.7) by $\tau_\epsilon$.

The following result generalizes Theorem 2 in [BQ22] from $\alpha = 1/2$ to $1/2 \leq \alpha \leq 1$, and is a step towards establishing the existence of a limiting fairness process. The proof is an adaptation of that in [BQ22] and is omitted here.

Proposition 3.5. For $\epsilon > 0$ and $1/2 \leq \alpha \leq 1$, the shifted fairness processes $\{S_{t} \eta_{\alpha,t}^{n}\}_{n \in \mathbb{N}}$ are tight under any non-idling policy.

[BQ22] observe that under most stationary policies the fairness processes converges on a faster scale than the scaled system length processes. Hence the limiting fairness process does not depend on $t$ after $\tau_0$ and is deterministic, i.e., $\eta_{t,\alpha} = \eta_{\alpha,t}$ for all $t > \tau_0$. For these cases, we refer to $\eta_{\alpha,t}$ as the limiting fairness measure, slightly abusing the terminology. The following theorem summarizes several findings of [BQ22] for various routing policies and generalized from $\alpha = 1/2$ to $1/2 \leq \alpha < 1$.

Theorem 3.6. Suppose $1/2 \leq \alpha < 1$ and let $\eta_{\alpha,t}^{SSF}$, $\eta_{\alpha,t}^{FSF}$, $\eta_{\alpha,t}^{LISF}$ and $\eta_{\alpha,t}^{RR}$ be the limiting fairness policies corresponding to longest-idle-server-first, uniformly random routing, fastest-server-first and slowest-server-first, respectively. Then,

$$
\eta_{\alpha}^{SSF}(\hat{A}) = \begin{cases} 
\delta_{\mu_{\max}(\hat{A})} & \text{if } t > \tau_0 \\
\delta_{\hat{A}} & \text{if } t \leq \tau_0
\end{cases}, \quad \eta_{\alpha}^{FSF}(\hat{A}) = \begin{cases} 
\delta_{\mu_{\min}(\hat{A})} & \text{if } t > \tau_0 \\
\delta_{\hat{A}} & \text{if } t \leq \tau_0
\end{cases} \quad \text{and}
$$

$$
\eta_{\alpha,t}^{LISF}(\hat{A}) = \eta_{\alpha,t}^{RR}(\hat{A}) = \begin{cases} 
\int_{\hat{A}} \mu F(dp) & \text{if } t > \tau_0 \\
\int_{\hat{A}} \mu (dp) & \text{if } t \leq \tau_0
\end{cases}.
$$

The proof of Theorem 3.6 for $1/2 < \alpha < 1$ follows the same steps in [BQ22] without any modification and is omitted here. However, similar results do not hold when $\alpha = 1$. Section 5 shows that the analysis for the suggested policies when $\alpha = 1$ is far more involved than in [BQ22].

4 Convergence of the Scaled System Length Processes

In this section, we derive fluid and diffusion approximations as the weak limit of the scaled system length processes $\{\hat{X}_{\alpha}^{n}\}_{n \in \mathbb{N}}$ when $1/2 \leq \alpha \leq 1$ using the fairness process defined in Section 3.2. Recall that the initial scaled system length processes $\hat{X}^{n}(0)$ is assumed to weakly converge to a non-positive random variable $\xi_0$. $\eta_{\alpha,t}$ is limiting fairness process under the routing policy employed by the system controller as in Definition 3.4 and $\langle t, \eta_{\alpha,t}\rangle$ corresponds to the first moment of the measure $\eta_{\alpha,t}$.
Theorem 4.1. Suppose that the limiting fairness process under the adopted routing policy is \( \eta_{\alpha, t} \). Then, the scaled system length processes \( \hat{X}_n^\alpha \Rightarrow \xi_\alpha \) as \( n \to \infty \), where

1. if \( \alpha = 1/2 \), \( \xi_\alpha \) is the strong solution to the stochastic differential equation

\[
\xi(t) = \xi_0 + \sqrt{2\lambda W(t)} - \left( \beta_F(\bar{\lambda}^\infty \mu_F^{1/2} + \zeta_1) \right) + \langle t, \eta_{1/2, t} \rangle \int_0^t (\xi(s))^{-\alpha} \, ds - \gamma \int_0^t (\xi(s))^{1+\alpha} \, ds, \tag{4.1}
\]

for all \( t \geq 0 \) where \( \zeta_1 \sim \text{Normal}(0, \sigma_F^2 \bar{\lambda}^\infty \mu_F^{1/2}) \) and \( W \) is a standard Brownian motion.

2. if \( 1/2 < \alpha \leq 1 \), the process \( \xi_\alpha \) is the solution to the ordinary differential equation

\[
\xi(t) = \xi_0 - \beta \bar{\lambda}^{\alpha, -\alpha} t + \langle t, \eta_{\alpha, t} \rangle \int_0^t (\xi(s))^{-\alpha} \, ds, \text{ for all } t \geq 0. \tag{4.2}
\]

A key question regarding the stochastic process limits above is whether it is possible to approximate the stationary behavior of the many-server queues using the diffusion and fluid limits given above, i.e., whether the many-server limit and the limit as \( t \to \infty \) is interchangeable. For example, for any \( 1/2 < \alpha \leq 1 \), if \( \eta_{\alpha, t} = \eta_{\alpha, \infty} \) for all \( t \geq \tau_0 \), then the solution to (4.2) can be written as

\[
\xi(t) = \frac{\beta \bar{\lambda}^{\alpha, -\alpha} t}{\langle t, \eta_{\alpha, \infty} \rangle} \left( \xi_0 + \frac{\beta \bar{\lambda}^{\alpha, -\alpha}}{\langle t, \eta_{\alpha, \infty} \rangle} \right) e^{-\langle t, \eta_{\alpha, \infty} \rangle t},
\]

and as \( t \to \infty \) we have \( \xi(t) \to \beta \bar{\lambda}^{\alpha, -\alpha} \langle t, \eta_{\alpha, \infty} \rangle^{-1} \). Using a similar reasoning, if \( \eta_{\alpha, t} \Rightarrow \eta_{\alpha, \infty} \), one can conclude that \( \xi(t) \to \beta \bar{\lambda}^{\alpha, -\alpha} \langle t, \eta_{\alpha, \infty} \rangle^{-1} \) as \( t \to \infty \). Thus, if the answer to the interchangeability question is affirmative, then for \( n \) large enough, we can approximate the long run number of idle servers as

\[
\mathbb{E}[\hat{I}_n^\alpha(\infty)] \approx \beta \bar{\lambda}^{\alpha, -\alpha} \langle t, \eta_{\alpha, \infty} \rangle^{-1}. \tag{4.3}
\]

This argument underpins much of the success mean-field games theory has collected in recent years, [BBC19] and references therein. As the next step, we justifying the approximation (4.3) by proving the interchangeability of limits in Theorem 4.2.

Theorem 4.2. For many-server systems with random and heterogeneous service rates, for any \( 1/2 \leq \alpha \leq 1 \) the following convergence results hold as \( n \to \infty \),

1. \( \hat{X}_n^\alpha(\infty) \Rightarrow \xi_\alpha(\infty) \)
2. \( \mathbb{E}[\hat{X}_n^\alpha(\infty)] \to \mathbb{E}[\xi_\alpha(\infty)] \)
3. \( \eta_n^\alpha \Rightarrow \eta_\infty \)

where \( \xi_\alpha(t) \Rightarrow \xi_\alpha(\infty) \) and \( \eta_t \Rightarrow \eta_\infty \) as \( t \to \infty \).

Theorems 4.1 and 4.2 demonstrate the importance of characterizing the limiting fairness process under a suggested routing policy. In Theorem 3.6 we presented already some results in this direction when \( 1/2 \leq \alpha < 1 \). In the next section, we concentrate on the purely-quality-driven regime and characterize the limiting fairness processes for a general class of randomized policies.
5 A Generalized Random Routing Policy for Purely-Quality-Driven Policies

In this section we address the following two questions: Given a service rate distribution \( F \), which limiting fairness processes are attainable? How should one design a routing policy to attain a specified fairness process? In Theorem 3.6 we have seen that for \( 1/2 \leq \alpha < 1 \), the limiting fairness process can be quite general and even measures with a mass concentrating at certain service rate values are attainable. However, when \( \alpha = 1 \), the dynamics differ substantially. To understand this difference, consider any \( A \in B(\mathbb{R}_+) \) with \( F(A) > 0 \). The number of agents with service rates in the set \( A \) scales with \( n \) and Theorem 4.1 implies that the total idleness observed in the system scales with \( n^\alpha \). Hence, when \( \alpha < 1 \), it is possible for all the idleness to concentrate on the agents of set \( A \) for \( n \) large enough. As a result, it is possible to have point masses in the limiting fairness measures as seen in Theorem 3.6. On the other hand, when \( \alpha = 1 \), i.e., in a purely quality driven regime, both the total idleness and the number of agents in \( A \) scale with \( n \), and the number of agents in \( A \) can potentially be less than the total idleness experienced in the system for any \( n \). As we always need to have the total idleness experienced by the agents in set \( A \) less than the number of agents in that set, we can quickly calculated for any \( n \in \mathbb{N} \) and \( T > 0 \),

\[
\frac{\sum_{k=1}^{N_n} \delta_{\tilde{\mu}_k^n}(A)}{T} \leq \sum_{k=1}^{N_n} \delta_{\tilde{\mu}_k^n}(A) \Rightarrow \frac{N_n \eta_{\alpha,\infty}(A)}{\sum_{k=1}^{N_n} \delta_{\tilde{\mu}_k^n}(A)} \leq \frac{N_n^\alpha T}{\sum_{k=1}^{N_n} \int_0^T I_k^n(t)dt}. \tag{5.1}
\]

When \( 1/2 \leq \alpha < 1 \), Theorem 4.1 implies that the right-hand side diverges to infinity as \( n \to \infty \), and hence, (5.1) does not pose any restriction on the limiting fairness measure. However, when \( \alpha = 1 \), taking the limit and replacing from (4.3), we get the following necessary condition for any probability measure to be the stationary limiting fairness measure under a routing policy.

**Theorem 5.1.** In the purely quality-driven regime, a necessary condition for a probability measure \( \eta \) to be the stationary limiting fairness measure under some routing policy is the

\[
0 \leq \frac{\eta_{\alpha,\infty}(A)}{F(A)} \leq \frac{(1 + \beta)\langle \eta \rangle}{\beta \tilde{\mu}_F}. \tag{5.2}
\]

Equation (5.2) implies that, in the purely quality-driven regime, the stationary limiting fairness measure is absolutely continuous and, hence, have a density \( g(\mu) \) with respect to \( F \). Considering sets \( [\mu, \mu + \Delta] \) and taking the limit as \( \Delta \to 0 \), (5.2) then takes the form

\[
0 \leq g(\mu) \leq \frac{(1 + \beta)\langle \eta \rangle}{\beta \tilde{\mu}_F}. \tag{5.3}
\]

We naturally ask the reverse implication, concretely, is the necessary condition provided in Theorem 5.1 also sufficient. In this section, we provide a generalized random routing policy that can attain any given stationary limiting fairness measure with a density that satisfies (5.3) strictly in the purely quality-driven regime. We first formally define our routing policy.

**Definition 5.2.** Given a real-valued function \( h(\mu) \) such that \( h(\mu) > 0 \) for all \( \mu_{\min} \leq \mu \leq \mu_{\max} \), a routing policy is called \( h \)-random if the probability of an incoming job arriving at time \( t \) to be routed upon arrival to the idle server \( k \) having service rate \( \tilde{\mu}_k^n \) is proportional to \( h(\tilde{\mu}_k^n) \) and is given by

\[
\frac{h(\tilde{\mu}_k^n)I_k^n(t-t^-)}{\sum_{i=1}^{N_n} h(\tilde{\mu}_i^n)I_i^n(t-t^-)},
\]

where \( 0/0 \) is interpreted as \( 0 \).
To be able to analyze $h$-random policies, we define the finite measure-valued instantaneous allocation of idleness processes as $\psi^n_t(\Lambda) = \sum_{k=1}^{N} \delta_{\bar{\pi}^k_t}(\Lambda) f_k^n(t)$, and this process denotes the number of servers with service rate in the set $\Lambda$ who are idle at time $t$ at the $n$th system. We also define the scaled instantaneous allocation as $\tilde{\psi}^n_t(\Lambda) = n^{-1} \psi^n_t(\Lambda)$.

**Theorem 5.5.** The set of scaled instantaneous allocations, $\{\tilde{\psi}^n_t\}_{n \in \mathbb{N}_0}$, is tight.

The above tightness result implies that any subsequence of $\tilde{\psi}^n_t$ has a further subsequence that converges. In Lemma 5.4, we provide an equation that the limits of all subsequences must satisfy.

**Lemma 5.4.** If the subsequence $\tilde{\psi}^{n_k}_t \Rightarrow \tilde{\psi}_t$ as $n_k \to \infty$. Then, $\tilde{\psi}_t$ satisfies
\[
\langle f, \tilde{\psi}_{1,t} \rangle = \langle f, \tilde{\psi}_{1,0} \rangle + \frac{\bar{\lambda}}{\bar{\mu}} (1 + \beta) (f \times \iota, F) \int_0^t \mathbb{I}(\xi_{\alpha,s} \leq 0) ds - \int_0^t (f \times \iota, \tilde{\psi}_{1,s-}) \mathbb{I}(\xi_{\alpha,s} \leq 0) ds \\
- \bar{\lambda} \int_0^t \frac{\langle f \times h, \tilde{\psi}_{1,s-} \rangle}{\langle h, \tilde{\psi}_{s-} \rangle} \mathbb{I}(\xi_{\alpha,s} \leq 0) ds \text{ for all } t \geq 0.
\] (5.4)

Equation (5.4) enables us to study the transient behavior of scaled instantaneous allocation processes. Theorem 4.2 implies that the stationary behavior of these processes also can be characterized by this equation. Our next result shows how these results can be used to obtain the stationary limiting fairness process.

**Theorem 5.5.** When $\alpha = 1$, the stationary limiting fairness measure $\eta_{1,\infty}$ under an $h$-random policy can be written as
\[
\eta_{1,\infty}(\Lambda) = \frac{\int_{\mu_{\min}}^{\mu_{\max}} (1 + L_F \bar{h}(\mu))^{-1} dF(\mu)}{\int_{\mu_{\min}}^{\mu_{\max}} (1 + L_F h(\mu))^{-1} dF(\mu)}, \quad \text{for all } \Lambda \in B[\mu_{\min}, \mu_{\max}],
\] (5.5)

where $\bar{h}(\mu) = h(\mu)/\mu$ and $L_F$ is the unique solution of
\[
\int_{\mu_{\min}}^{\mu_{\max}} \frac{1 + \beta}{\bar{\mu}_F (1 + L_F \bar{h}(\mu))} dF(\mu) = \beta.
\] (5.6)

The following lemma provides bounds on $L_F$, the solution of (5.6).

**Lemma 5.6.** Define $h_{\min} = \min_{\mu_{\min} \leq \mu \leq \mu_{\max}} h(\mu)$ and $h_{\max} = \max_{\mu_{\min} \leq \mu \leq \mu_{\max}} h(\mu)$ and let $L_F$ be the solution of (5.6) for some service rate distribution $F$. Then,
\[
\frac{\beta}{h_{\min}} \leq L_F \leq \frac{\beta}{h_{\max}}.
\] (5.7)

Equation (5.5) implies that the density of the stationary limiting fairness process with respect to the service rate distribution $F$ is
\[
g(\mu) = \frac{(1 + L_F \bar{h}(\mu))^{-1}}{\int_{\mu_{\min}}^{\mu_{\max}} (1 + L_F \bar{h}(\mu))^{-1} dF(\mu)}.
\]

At the beginning of this section, we provided the necessary condition (3.3) stating that for a function $g(\mu)$ to be the density function of $\eta_{1,\infty}$ with respect to $F$, it needs to satisfy
\[
0 \leq g(\mu) \leq \frac{(1 + \beta)(\mathcal{L}, \eta_{1,\infty})}{\beta \bar{\mu}_F}.
\] (5.8)

Our next result shows that it is possible to find a suitable $h$-random policy to attain any density which satisfies the inequalities in (5.8) strictly.
Corollary 5.7. Suppose that $g(\mu)$ satisfies (5.8) strictly. Then, $g(\mu)$ is the density function of $\eta_{1,\infty}$ with respect to $F$ under the $h$-random policy where

$$h(\mu) = \frac{(1+\beta \bar{\mu} F - \beta \langle \iota, \eta_{1,\infty} \rangle)^{-1} g(\mu)}{\beta \lambda(\iota, \eta_{1,\infty})^{-1} g(\mu)}.$$ 

6 Strategic Servers with Individual Preferences

In this section, we deploy the results of the previous sections to analyze the strategic behavior of servers in a many-server setting when each server has individual preferences as described in Section 3. As the first step, we show how the stationary limiting fairness process can be used to characterize the idleness experienced by individual servers. This again requires an interchange of limits argument and we show that this interchange is valid for $h$-random policies under purely-quality driven staffing ($\alpha = 1$) and for any idle-time order based policy when $1/2 \leq \alpha \leq 1$. Then, we study the best response of server $k$ with characteristics determined by $\tilde{\mu}_n^{\min, k}, \tilde{\mu}_n^{\max, k}$, and then use our analysis to derive a fixed point equation that characterizes the Nash equilibria of the strategic servers game. We have the following assumptions on server characteristics.

Assumption 6.1. 1. The trade-off coefficient $\tilde{a}_n^k$ is bounded away from zero and bounded from above, i.e., there exists $a_{\min} > 0$ and $a_{\max} < \infty$ such that $P(a_{\min} \leq \tilde{a}_n^k < a_{\max}) = 1$.

2. The derivative of the cost of effort function is bounded away from zero and bounded from above, i.e., there exists $c_{\min} > 0$ and $c_{\max} < \infty$ such that $c_{\min} \leq c'(\mu) \leq c_{\max}$ for all $\mu_{\min} \leq \mu \leq \mu_{\max}$.

3. The limiting stationary fairness measure is deterministic and absolutely continuous with respect to $F$ with a continuously differentiable density function $g(\mu)$ on the closed interval $[\mu_{\min}, \mu_{\max}]$. Hence, both $g(\mu)$ and its derivative $g'(\mu)$ are bounded from above, i.e., $g(\mu) \leq g_{\max}$ and $g'(\mu) \leq g'_{\max}$, and $g'(\mu)$ is bounded away from 0, i.e., $0 < g'_{\min} \leq g'(\mu)$ for all $\mu_{\min} \leq \mu \leq \mu_{\max}$.

The last assumption on the limiting stationary fairness measure being deterministic might seem restrictive. However, as we have seen in Theorems 3.6 and 5.5, this property holds under all routing policies considered in this work.

6.1 The Expected Long Run Proportion of Idleness Experienced by Individual Servers

To relate the stationary limiting fairness measure to the idleness experienced by individual servers, we need to make sure that the routing policy is indifferent to the index of the servers, and if two servers choose to serve with the same rate their expected long run proportion of idleness should be equal, i.e.,

$$E[I_n^k(\infty) | \hat{\mu}_n^k = \mu] = E[I_n^j(\infty) | \hat{\mu}_n^j = \mu] \text{ for all } 1 \leq k, j \leq N_\alpha \text{ and } \mu_{\min} \leq \mu \leq \mu_{\max}. \quad (6.1)$$

Then, defining the empirical law of service rates for the $n$th system as

$$F_n^\alpha[\mu, \mu + \Delta] = \frac{\sum_{j=1}^{N_\alpha} \delta_{\mu_n^{j}}[\mu, \mu + \Delta]}{N_\alpha^\alpha} \text{ for all } \mu \in [\mu_{\min}, \mu_{\max}] \text{ and } \Delta > 0,$$
then for any \( n \in \mathbb{N} \) can calculate
\[
\mathbb{E}[n^{1-\alpha} I_k^n(\infty) | \mu_k^n = \mu] = \lim_{\Delta \to 0} \mathbb{E} \left[ n^{1-\alpha} \sum_{j=1}^{N_{\alpha}} \delta_{\mu_j^n}([\mu_j^n + \Delta]) I_j^n(\infty) | \mu_k^n = \mu \right]
\]
\[
= \lim_{\Delta \to 0} \mathbb{E} \left[ \frac{\sum_{j=1}^{N_{\alpha}} \delta_{\mu_j^n}([\mu_j^n + \Delta]) I_j^n(\infty)}{\bar{F}_{\alpha}(\infty)} \bar{F}_{\alpha}(\infty) \frac{n}{\sum_{j=1}^{N_{\alpha}} \delta_{\mu_j^n}([\mu_j^n + \Delta]) | \mu_k^n = \mu} \right]
\]
\[
= \lim_{\Delta \to 0} \mathbb{E} \left[ \frac{n^{1-\alpha} \delta_{\mu_j^n}([\mu_j^n + \Delta]) I_j^n(\infty) \frac{n}{\sum_{j=1}^{N_{\alpha}} \delta_{\mu_j^n}([\mu_j^n + \Delta]) | \mu_k^n = \mu}}{F^n([\mu_j^n + \Delta]) \bar{F}_{\alpha}(\infty) \sum_{j=1}^{N_{\alpha}} \delta_{\mu_j^n}([\mu_j^n + \Delta]) | \mu_k^n = \mu} \right].
\]
Taking the limit as \( n \to \infty \) yields
\[
\lim_{n \to \infty} \lim_{\Delta \to 0} \mathbb{E}[n^{1-\alpha} I_k^n(\infty) | \mu_k^n = \mu] = \lim_{n \to \infty} \lim_{\Delta \to 0} \mathbb{E} \left[ \frac{n^{1-\alpha} \delta_{\mu_j^n}([\mu_j^n + \Delta]) I_j^n(\infty) \frac{n}{\sum_{j=1}^{N_{\alpha}} \delta_{\mu_j^n}([\mu_j^n + \Delta]) | \mu_k^n = \mu}}{F^n([\mu_j^n + \Delta]) \bar{F}_{\alpha}(\infty) \sum_{j=1}^{N_{\alpha}} \delta_{\mu_j^n}([\mu_j^n + \Delta]) | \mu_k^n = \mu} \right].
\]
Equation (6.2) implies that the idleness received by a server is in the order of \( n^{\alpha-1} \) and approaches 0 if \( \alpha < 1 \). When the utility of idleness function is identity, i.e., it is identically equal to the expected long run proportion of idleness \( x \), as in [GDWW16, Gop21, ARS21], setting \( \alpha < 1 \) does not provide enough incentive for the servers to work harder (increase their service rate) and results in all servers working with their minimal service rate \( \bar{\mu}_{\min,k} \) for large \( n \). Hence, as [GDWW16] proves, for servers with identical preferences, purely quality-driven regime is asymptotically optimal in this case.

When the utility of idleness function is the identity function, it is non-negative and the utility is zero when the expected long run proportion of idleness is zero. Even though this is good for modeling the benefits of idleness, it fails to appropriately model the discomfort due to the server not experiencing enough idleness. Similarly, when the utility of idleness function is bounded from below, it can be made equivalent to a non-negative utility of idleness by adding a constant. Hence, to model the situations where servers are sensitive to experiencing a low level of idleness and experiencing no idleness is unacceptable by the servers, one needs to have \( f(x) \to \infty \) as \( x \to 0 \). In this case, the servers might be inclined to work faster, even though the proportion of idleness experienced by the server is very close to zero, to reduce the discomfort. This suggests that if the servers are sufficiently sensitive to experiencing a low level of idleness, purely quality-driven regime is suboptimal – this cannot be recovered from [GDWW16] since there \( f \) is the identity function. Under the assumption that the convergence in (6.4) is uniform, Theorem 6.2 shows how the safety staffing level and the rate which \( f(\cdot) \) approaches infinity as idleness goes to 0 determines the best response rate of a server.
Theorem 6.2. If $\sup_{\mu_{\min} \leq \mu \leq \mu_{\max}} |n^{1-\alpha}E[\Pi^n_k(\infty)|\bar{\mu}^n_k = \mu] - \beta g_F(\mu)\bar{\lambda}^{-1}\bar{\mu}^{2-\alpha}(t, \eta_{\alpha, \infty})^{-1}| \to 0$, we have

1. if $n^{\alpha-1}f'(n^{\alpha-1}x) \to 0$ for all $x > 0$ as $n \to \infty$, then for any $\epsilon > 0$ there exists an $N_\epsilon$ such that $n > N_\epsilon$ implies that the optimal strategy for any server $k$ in the $n$th system is $\tilde{\mu}^{n,k}_k \in [\bar{\mu}^n_{\min,k}, \bar{\mu}^n_{\min,k} + \epsilon]$.

2. if the limiting stationary fairness measure is deterministic with a decreasing density $g_F(\mu)$ for any service rate distribution $F$, then for any $\epsilon > 0$ there exists an $N_\epsilon$ such that $n > N_\epsilon$ implies that the optimal strategy for any server $k$ in the $n$th system is $\tilde{\mu}^{n,k}_k \in [\bar{\mu}^n_{\min,k}, \bar{\mu}^n_{\min,k} + \epsilon]$.

3. if $n^{\alpha-1}f'(n^{\alpha-1}x) \to \infty$ for all $x > 0$ as $n \to \infty$ and $g_F(\mu)$ is strictly increasing and concave for any service rate distribution $F$, then for any $\epsilon > 0$ there exists an $N_\epsilon$ such that $n > N_\epsilon$ implies that the optimal strategy for any server $k$ in the $n$th system is $\tilde{\mu}^{n,k}_k \in (\tilde{\mu}^n_{\max,k} - \epsilon, \bar{\mu}^n_{\min,k}]$ with probability 1.

Theorem 6.2 has clear implications regarding the optimal scaling for the safety staffing. When $\alpha < 1$, the dominating term in the total staffing cost (where $c_S$ is the unit staffing cost)

$$c_S \frac{\lambda^n}{\bar{\mu}_F} + c_S \left(\frac{\lambda^n}{\bar{\mu}_F}\right)^\alpha,$$

is due to the offered load. Hence, for large $n$, the system operator mainly aims to maximize the average service rate in order to minimize the staffing cost. Suppose that for $\alpha_0 < 1$, we have $n^{\alpha_0-1}f'(n^{\alpha_0-1}x) \to f'(x) < \infty$ which is not identically zero. Then, setting $\alpha_0 < \alpha < 1$ and adopting a policy which results in an increasing concave density $g(\mu)$ achieves the maximum possible expected service rate $\bar{\mu}_F = E[\tilde{\mu}^n_{\max,k}]$.

As we have seen, the interchange of limits in (6.3) has implications leading to Theorem 6.2. By definition, the stationary fairness measure $\eta^{n, \infty}_n$ is absolutely continuous, and hence, has a density $g^n(\cdot)$ with respect to $F^n$ for all $n \in \mathbb{N}$ and (6.3) is equivalent to $g^n(\mu) \to g(\mu)$ as $n \to \infty$. In general, weak convergence of measures does not necessarily imply the convergence of densities. However, we believe that (6.3) holds for most of the reasonable routing policies. Below, we show that (6.3) holds for some of the specific classes of policies we discussed in the previous sections. In particular, the following theorem shows that the limits can be interchanged for any $h$-random policy under a purely quality driven regime.

Theorem 6.3. For $\alpha = 1$, under an $h$-random routing policy, Equation (6.3) holds and the convergence is uniform, i.e.,

$$\sup_{\mu_{\min} \leq \mu \leq \mu_{\max}} \left| E[I^n_k(\infty)|\bar{\mu}^n_k = \mu] - (1 + L_F h(\mu))^{-1} \right| \to 0.$$

[GDWW16] introduced the class of idle-time order based policies, namely, the class of policies where the selected server to which an arrival is routed to depends only on the order in which the servers last became idle. Some common policies such as longest-idle-server-first and random routing are also in this class. We refer the reader to [GDWW16] for a formal definition. Using Theorem 9 in [GDWW16], our next theorem shows that (6.3) holds for idle-time order based policies under quality-and-efficiency driven regime as well as any quality-driven regime.

Theorem 6.4. For $1/2 \leq \alpha \leq 1$, under any idle-time order based policy, (6.3) holds and

$$\sup_{\mu_{\min} \leq \mu \leq \mu_{\max}} \left| n^{1-\alpha}E[I^n_k(\infty)|\bar{\mu}^n_k = \mu] - \mu \beta \bar{\lambda}^{-1}(\bar{\mu}^n_F \sigma_F^2 + \bar{\mu}^{2-\alpha}_F) \right| \to 0, \quad for \ all \ 1/2 \leq \alpha < 1.$$
and

\[
\sup_{\mu_{\min} \leq \mu \leq \mu_{\max}} \left| E[T_k^n(\infty)|\tilde{\mu}_k^n = \mu] - \frac{\beta(t, \eta_{1,\infty})^{-1}}{\mu\beta(t, \eta_{1,\infty})^{-1} + 1} \right| \to 0, \quad \text{for } \alpha = 1,
\]
as \(n \to \infty\), where \(\eta_{1,\infty}\) is the limiting fairness measure for the \(h\)-random policy with \(h(\mu) = 1\) as given in Theorem 5.5.

### 6.2 The best response of a server, and the characterization of Nash equilibria under an \(h\)-Random policy in purely-quality driven regime \((\alpha = 1)\)

In this section, our goal is to characterize the Nash equilibria when the servers are strategic with heterogeneous abilities and preferences under a given \(h\)-random policy in the limiting system. To do so, we start by analyzing the best response strategy of a given server \(k\) with parameters \(\tilde{\mu}_{\min,k}, \tilde{\mu}_{\max,k}\) and \(\tilde{a}_k\), when the distribution of service rate for all other servers is \(F_0\) (i.e., known). Using Theorem 6.3, the utility maximization for server \(k\) in the \(n\)-limit takes the form

\[
U_k^* := \max_{\tilde{\mu}_{\min,k} \leq \mu \leq \tilde{\mu}_{\max,k}} U_k(\mu) = \max_{\tilde{\mu}_{\min,k} \leq \mu \leq \tilde{\mu}_{\max,k}} f\left((1 + L_{F_0}\tilde{h}(\mu))^{-1}\right) - \tilde{a}_k c(\mu).
\]

It is clear that when \(\tilde{h}(\mu)\) is non-decreasing, the idleness observed by server \(k\) will be non-increasing and the best response of the server in this case will be to set her service rate to \(\mu_k = \tilde{\mu}_{\min,k}\). Also, as standard in the literature, we need concavity properties for the utility function. Hence, we have the following assumption:

**Assumption 6.5.** The function \(\tilde{h}(\mu)\) is convex strictly decreasing satisfying

\[
2\tilde{h}'(\mu)^2 \leq \tilde{h}(\mu)\tilde{h}''(\mu), \quad \text{for all } \mu_{\min} \leq \mu \leq \mu_{\max}.
\]

Condition (6.7) is needed to ensure concavity of the limiting utility function without making additional structural assumptions on the utility of idleness function \(f(\cdot)\). It is satisfied by any \(\tilde{h}(\mu) = \mu^{-p}\), where \(0 < p \leq 1\), which have an interesting managerial interpretation. The case \(p = 1\) implies \(h(\mu) = 1\) and corresponds to idle-time order based policies, introduced by [GDWW16], which routes arrivals to servers without taking their service rates into account. When \(p < 1\), \(h(\mu) = \mu^{-p}\) and the system operator is more eager to route arrivals to servers with high service rates, but just so that the idleness a server receives is still an increasing function of the service rate. If we have additional information on \(f(\cdot)\), it is possible to obtain convexity without needing (6.7), e.g., if \(f(x) = -x^{-1}\), then the utility function is concave for any convex \(h(\mu)\). For these cases, the results obtained in this section still hold. The next lemma shows that the limiting utility function is concave under Assumption 6.5.

**Lemma 6.6.** Under Assumption 6.5 the limiting utility function \(U_k(\mu)\) is concave for any fixed \(L_{F_0}\).

Now, as Lemma 6.6 ensures that the second order optimality conditions are satisfied, we can concentrate on the first order conditions. Taking the derivative of the utility function and after some algebraic manipulations, first order conditions take the form

\[
C(\mu, L_{F_0}) = -\frac{L_{F_0}f'\left((1 + L_{F_0}\tilde{h}(\mu))^{-1}\right)\tilde{h}'(\mu)}{(1 + L_{F_0}\tilde{h}(\mu))\tilde{h}''(\mu)} = \tilde{a}_k.
\]
Due to our convexity/concavity assumptions, the left-hand side of (6.8) is decreasing in \( \mu \) and its solutions maximize the utility of a server. Based on Assumption 6.5, letting \( \mu_k^* \) be the smallest of these solutions, the best response of server \( k \), \( \mu_k^* \), is given by

\[
\mu_k^* = \begin{cases} 
\hat{\mu}_{\text{min},k}, & \text{if } \hat{a}_k \geq C(\hat{\mu}_{\text{min},k}, L_{F_0}) \\
\mu_k^*, & \text{if } C(\hat{\mu}_{\text{min},k}, L) \leq \hat{a}_k \leq C(\hat{\mu}_{\text{max},k}, L_{F_0}) \ \\
\hat{\mu}_{\text{max},k}, & \text{if } \hat{a}_k \leq C(\hat{\mu}_{\text{max},k}, L_{F_0}) 
\end{cases}
\]  

(6.9)

Considering that the parameters \( \hat{\mu}_{\text{min},k}, \hat{\mu}_{\text{max},k} \) and \( \hat{a}_k \) are random with respective distributions described in Section 3.1, we obtain that the distribution function of the optimal service rate of any server \( k \) is

\[
F_1(\mu | L_{F_0}) := \mathbb{P}(\mu_k^* \leq \mu) = \mathbb{P}(\hat{\mu}_{\text{min},k} \leq \mu) + \mathbb{P}(\hat{\mu}_{\text{max},k} > \mu, \hat{\mu}_{\text{min},k} < \mu, \hat{a}_k \geq C(\mu, L_{F_0})).
\]  

(6.10)

We are thus able to characterize the equilibrium service rate distribution as follows.

**Definition 6.7.** Let \( F_0 \) be a distribution function with support \([\mu_{\text{min}}, \mu_{\text{max}}]\). We call \( F_0 \) an equilibrium service rate distribution if \( F_0(\mu) = F_1(\mu | L_{F_0}) \) is as given in (6.10) for all \( \mu \in [\mu_{\text{min}}, \mu_{\text{max}}] \).

Note that the best response of server \( k \) depends on \( F_0 \) only through \( L_{F_0} \), and \( L_{F_0} \) is uniquely determined by \( F_0 \) through (5.6). This \( L_{F_0} \) is then used to build a new distribution \( F_1 \) in the form given in (6.10) for the service rates, and this in turn yields a new \( L_{F_1} \). This process defines an operator \( \mathcal{L} \) which maps \( L_{F_0} \) to \( L_{F_1} \), i.e., \( \mathcal{L}(L_{F_0}) = L_{F_1} \). Hence, the fixed points of the mapping \( \mathcal{L} \), i.e., the set \( \{ L_F : \mathcal{L}(L_F) = L_F \} \), characterize the Nash equilibria. In other words, the equilibrium service rate distribution should have the form \( F(\mu | L_F) \) where the solution of (5.6) is also \( L_F \). The next theorem summarizes these arguments.

**Theorem 6.8.** If \( F \) is an equilibrium service rate distribution, then the distribution function \( F \) has the form given in (6.10) where \( L_F \) is the solution of

\[
\int_{\mu_{\text{min}}}^{\mu_{\text{max}}} \frac{1 - \beta L_F \hat{h}(\mu)}{1 + L_F \hat{h}(\mu)} dF(\mu | L_F) = 0.
\]  

(6.11)

Under the additional assumption that \( F_a \) is a continuous distribution, the next result shows that a solution to Equation (6.11) exists and is in the interval provided in Lemma 5.6. Moreover, all solutions lie in this interval.

**Proposition 6.9.** Let the distribution of \( \hat{a}_k^n, F_a \) be continuous. Then, Equation (6.11) has at least a solution in the interval \([1/(\beta \hat{h}(\mu_{\text{min}})), 1/(\beta \hat{h}(\mu_{\text{max}}))]\) and has no solution outside of this interval.

### 6.3 Numerical Experiments

In this section, we numerically analyze the equilibria to gain further insight into the problem. For simplicity, we assume that the vector \((\hat{\mu}_{\text{min},k}, \hat{\mu}_{\text{max},k})\) is independent of \( \hat{a}_k \). We assume that these have bounded support with \( \mu_{\text{min}}^k \leq \hat{\mu}_{\text{min},k} < \hat{\mu}_{\text{max},k} \leq \mu_{\text{max}}^k \) with the density functions \( f_{\text{min},\text{max}}(\mu_1, \mu_2) = 2(\mu_{\text{max}} - \mu_{\text{min}})^{-2} \) and \( f_a(a) = a_{\text{max}} - a_{\text{min}} \). We also assume that the utility of idleness, cost of effort and the routing functions are power functions, i.e., \( f(\mu) = \mu^p, c(\mu) = \mu^q \) and \( h(\mu) = \mu^r \). Table 1 presents the data we use as the base-case of our numerical experiments.

Proposition 6.9 shows that equation (6.11) has a solution, possibly non-unique, in the interval suggested by Lemma 5.6. We start by investigating the uniqueness of a solution to (6.11) by plotting...
The integral value as function of $L_F$ on said interval for various values of $\beta$ and $r$. Figure 6.2b illustrates the value of the integral on the left-hand side of (6.11) as a function of $L_F$, and shows all these functions to be equal to 0 for a unique value of $L_F$ for each fixed $r$ and $\beta$. We performed the same experiments for different values of $p$ and $q$ and obtained similar graphs (which we do not present).

Figure 6.1: Integral on the left-hand side of (6.11) as function of $L_F$ for various values of $r$ and $\beta$.

![Figure 6.1](image)

Figure 6.2 displays how the routing policy affects the distribution of $\tilde{\mu}_k$ and the staffing level $(3.1)$ by varying $\mu = \mu'$, Figure 6.2(a) shows that the equilibrium $L_F$, the solution of (6.11), is increasing in $r$. Using these equilibrium values and the base-case data, we determine the distribution $\tilde{\mu}_F$ and staffing level $N$. As $r$ increases, it is more likely for an arrival to be routed to a server having high service rate compared to servers with lower service rates. Figure 6.2(b) shows that as $r$ increases both the incentive for servers to work faster and $\tilde{\mu}_F$ decreases. Consequently, the staffing level $N$ increases as $r$ increases for a fixed $\beta$, see Fig. 6.2(c). Finally, Figure 6.2(d) illustrates the density function $f(\mu|L_F)$ of service rates for various values of $r$. For $\mu$ very close to 0, we observe that $C(\mu, L_F)$ is very high and hence the density function is very close to $f_{\text{max}}$. Due to $a$ having bounded support, we see a spike in the density function when $C(\mu, L_F) = a_{\text{max}}$ and after that point the density decreases as $\mu$ increases.

Finally, we numerically explore how the safety-staffing coefficient $\beta$ affects the system. Similar to above, we see that $L_F$ decreases (Fig. 6.3(a)) and $\tilde{\mu}_F$ increases as $\beta$ increases (Fig. 6.3(b)). Also, the distribution service rates exhibit the same spikes (as when Fig. 6.2(d)) at $C(\mu, L_F) = a_{\text{max}}$ (Fig. 6.3(d)). However, Fig. 6.3(c) shows that an optimal level of staffing exists around $\beta = 0.4$.

---

### Table 1: The base-case scenario for experiments.

| $\lambda$ | $\beta$ | $\mu_{\text{min}}$ | $\mu_{\text{max}}$ | $a_{\text{min}}$ | $a_{\text{max}}$ | $p$ | $q$ | $r$ |
|-----------|---------|---------------------|---------------------|------------------|------------------|-----|-----|-----|
| 100       | 0.3     | 0.01                | 0.5                 | 0.01             | 25               | 1   | 2   | -1 |
Figure 6.2: The effects of the routing policy on $F(\mu | L_F)$ and $N$.

(a) $r$ vs. $L_F$

(b) $r$ vs. $\mu_F$

(c) $r$ vs. $N$

(d) $\mu$ vs. $F(\mu | L_F)$ in $r$
Figure 6.3: The effects of the safety staffing level $\beta$ on $F(\mu|L_F)$ and $N$.

(a) $\beta$ vs. $L_F$

(b) $\beta$ vs. $\mu_F$

(c) $\beta$ vs. $N$

(d) $\mu$ vs. $F(\mu|L_F)$ in $\beta$
7 Concluding Remarks

The recent evidence from call centers suggest that servers exhibit a call avoidance behavior when they strategically adjust their service rates in order to experience a certain level of idleness. Recently, there has been some research effort to model this strategic behavior to understand how routing in service systems affect the system performance. Due to the complexity of the problem, this line of research mainly has concentrated on identifying symmetric equilibrium assuming that all servers are identical. However, due to the servers being humans, they have inherently different preferences and capabilities. In this work, we introduce a method to study strategic behavior of servers when servers have heterogeneous abilities and preferences.

Our method relies on the concept of fairness process introduced by [BQ22] which describes how the idleness is distributed among servers with different rates. In this work, we first use the fairness processes to prove fluid limits for many-server service systems under quality-driven regimes. Then, we identify the limiting distribution of idleness in stationarity under a generalized random routing policy, where the probability that an incoming customer is routed to a specific server depends on the service rate. We characterize the Nash equilibria in this generality and provide numerical examples showing how this characterization can be used to analyze systems with heterogeneous strategic servers.

In addition to modeling server heterogeneity and general random routing policies, we extend the literature on strategic servers to capture the discomfort due to experiencing low levels of idleness. To do so, we model the utility function as a concave increasing function of the experienced idleness. For the case where utility of idleness is modeled as idleness itself, [GDWW16] show that purely the quality-driven regime is asymptotically optimal. However, we show that if the discomfort increases fast enough as the idleness approaches zero, the quality-and-efficiency-driven regime and other quality driven regimes can be optimal. This is the first time this conclusion appears in the literature.

Our results show that the key to analyze systems with heterogeneous servers is to characterize how idleness is distributed among servers under a given routing policy, which is expressed in the limiting fairness process. We presented closed-form expressions for the limiting fairness processes for some priority-based policies such as fastest- and slowest-server-first under quality-and-efficiency regime and some quality-driven regimes. However, the characterization of the limiting fairness process for these routing policies under purely quality-driven regime still remains an open problem. Similarly, in this work, we characterize the limiting fairness process for h-random policies under purely-quality driven regime, but unfortunately, the same techniques do not work for other regimes.

To the best of our knowledge, all the literature on queues with strategic servers, including this work, focuses on analyzing equilibria with a single queue and single pool of servers. In most service systems, both the customers and the servers are classified into different queues and pools depending on their needs and abilities. We believe a particularly interesting and challenging research direction is the analysis of strategic server behavior under skill-based routing.

A Proofs for Theorems

A.1 Additional Notation

In addition to the notation described in Section 1.2, we define the short-hand notation $x \wedge y := \min\{x, y\}$. Set $d_S$ to be the metric for Skorokhod-$J_1$ topology and lastly, we denote the optional quadratic variation of a stochastic process $Y(t)$ as $[Y]_t$. 

21
A.2 Proofs for Results Presented in Section 3.2

Proof of Lemma 3.3 Let $T > 0$ and $n \in \mathbb{N}$, and define $\theta^{n,0} := \inf \{t \geq 0 : (\hat{X}_n^\alpha(t))^+ \leq 0 \}$. Then, for all $0 \leq t \leq \theta^{n,0}$,

\[
(\hat{X}_n^\alpha(t))^+ \leq (\hat{X}_n^\alpha(0))^+ + n^{-\alpha} A^n(t) - n^{-\alpha} \sum_{k=1}^{N_n} S_k(\tilde{\mu}_k^n t),
\]

\[
\leq (\hat{X}_n^\alpha(0))^+ + n^{-\alpha} (A^n(t) - \lambda^n t) - n^{-\alpha} \sum_{k=1}^{N_n} (S_k(\tilde{\mu}_k^n t) - \tilde{\mu}_k^n t)
\]

\[ + n^{-\alpha} (\lambda^n - N_n^n \bar{\mu}_F) t - n^{-\alpha} \sum_{k=1}^{N_n} (\tilde{\mu}_k^n - \bar{\mu}_F) t \]

\[ \xrightarrow{P} \zeta_0 - \beta \bar{\lambda}\bar{\mu}_F \bar{\alpha} t. \quad (A.1) \]

Defining $\theta^{n,1}_c := \inf \{s > \theta^{n,0} : (\hat{X}_n^\alpha(s))^+ > \epsilon \}$ and $\tilde{\theta}^{n,1}_c := \sup \{\theta^{n,0} \leq s < \theta^{n,1}_c : (\hat{X}_n^\alpha(s))^+ < \epsilon / 2\}$,

\[
\mathbb{P} \left( \sup_{\theta^{n,0} \leq s \leq T} (\hat{X}_n^\alpha(s))^+ > \epsilon \right) \leq \mathbb{P} \left( \theta^{n,0} < \tilde{\theta}^{n,1}_c \leq \theta^{n,1}_c \leq T \right)
\]

\[ \leq \mathbb{P} \left( \sup_{\theta^{n,0} \leq s \leq T} \left\{ \frac{|A^n(s_2) - A^n(s_1) - \lambda^n(s_2 - s_1)|}{n^\alpha} \right. \right.
\]

\[ + \left. \left. \sum_{k=1}^{N_n} \left( S_k(\tilde{\mu}_k s_1) - S_k(\tilde{\mu}_k s_2) - \tilde{\mu}_k(s_2 - s_1) \right) \right\} > \epsilon / 2 \right) \]

\[ \leq 2\mathbb{P} \left( n^{-\alpha} |A^n(s) - \lambda^n s_1^+ T > \epsilon / 16 \right) + 2\mathbb{P} \left( n^{-\alpha} \sum_{k=1}^{N_n} |S_k^n(s) - \tilde{\mu}_k^n| T > \epsilon / 16 \right) \]

\[ + \mathbb{P} \left( n^{-\alpha} \left( \lambda^n - \sum_{k=1}^{N_n} \tilde{\mu}_k^n \right) > 0 \right). \quad (A.2) \]

The right-hand side of (A.2) converges 0 and combining (A.1) and (A.2) we obtain stochastic boundedness of \{(X_n^\alpha(t))^+\}_{n \in \mathbb{N}}. Now, setting $t_M = (M_0 + 1) / \bar{\lambda}^\alpha \bar{\mu}_F^{-\alpha}$

\[
\mathbb{P} \left( \sup_{T M \leq s \leq T} (\hat{X}_n^\alpha(s))^+ > \epsilon \right) \leq \mathbb{P} \left( \sup_{\theta^{n,0} \leq s \leq T} (\hat{X}_n^\alpha(s))^+ > \epsilon, \theta^{n,0} \leq t_M \right) + \mathbb{P} (\theta^{n,0} > t_M)
\]

Equation (A.2) implies that the first probability on the right-hand side converges to 0. For the second term, $\theta^{n,0} > t_M$ implies $(\hat{X}_n^\alpha(t_M))^+ > 0$. Equation (A.1) implies that the probability of this event approaches 0, which proves the second part of the lemma.
We now prove the stochastic boundedness of \( \{ \hat{I}_\alpha^n(t) \}_{n \in \mathbb{N}} \). The case when \( \alpha = 1/2 \) was established in Lemma 1 in [BQ22]. The proof for \( 1/2 < \alpha \leq 1 \) is simpler. We have

\[
\left| \hat{I}_\alpha^n(t) \right|_T^* \leq \left| \hat{X}_\alpha^n(0) \right| + \left| \frac{A^n(t) - \lambda^n t}{n^\alpha} \right|_T^* + \left| \sum_{k=1}^{N^n_k} S^n_k(t) - \mu_k^n t \right|_T^* + \left| \sum_{k=1}^{N^n_k} \mu_k^n t - \lambda^n t \right|_T^* \\
+ \left| R^n \left( \gamma \int_0^t (X^n(s) - N^n)^+ ds \right) - \gamma \int_0^t (\hat{X}_\alpha^n(s))^+ ds \right|_T^* + \left| \gamma \int_0^t (\hat{X}_\alpha^n(s))^+ ds \right|_T^*.
\]

The stochastic boundedness of \( \{ (\hat{X}_\alpha^n(t))^+ \}_{n \in \mathbb{N}} \) and the martingale central limit theorem implies that the last two terms converge to 0 in probability. The second and third terms converge to 0 in probability due to law of large numbers for Poisson processes. Finally, the first and the fourth terms are stochastically bounded due to Assumption [3.2] and the central limit theorem. This proves the lemma. \( \square \)

A.3 Proofs for Results Presented in Section 4

**Proof of Theorem 4.1** Our proof follows the same steps as Theorem 5 in [BQ22] with the additional modification on the number of servers. Adding and subtracting appropriate terms to (3.4) and normalizing with \( n^\alpha \), we get

\[
\hat{X}_\alpha^n(t) = \hat{X}_\alpha^n(0) + \hat{M}_{\alpha,1}^n(t) - \hat{M}_{\alpha,2}^n(t) - \hat{M}_{\alpha,3}^n(t) + n^{-\alpha} (\lambda^n - N^n_{\alpha} \bar{\mu}_F) t - n^{-\alpha} \left( \sum_{k=1}^{N^n_k} \mu_k^n - N^n_{\alpha} \bar{\mu}_F \right) t
\]

\[
= - \sum_{k=1}^{N^n_k} \mu_k^n \int_0^t \hat{I}_{k,\alpha}^n(s) ds - \gamma \int_0^t (\hat{X}_\alpha^n(s))^+ ds, \quad (A.3)
\]

where

\[
\hat{M}_{\alpha,1}^n(t) := \frac{A^n(t) - \lambda^n t}{n^\alpha}, \\
\hat{M}_{\alpha,2}^n(t) := \frac{S^n \left( \sum_{k=1}^{N^n_k} \hat{\mu}_k^n \left( t - \int_0^t \hat{I}_k^n(s) ds \right) \right) - \sum_{k=1}^{N^n_k} \mu_k^n \left( t - \int_0^t \hat{I}_k^n(s) ds \right)}{n^\alpha}, \\
\hat{M}_{\alpha,3}^n(t) := \frac{R^n \left( \gamma \int_0^t (X^n(s) - N^n_{\alpha})^+ ds \right) - \gamma \int_0^t (\hat{X}_\alpha^n(s))^+ ds}{n^\alpha}.
\]

Using martingale central limit theorem, both \( \hat{M}_{\alpha,1}^n(t) \) and \( \hat{M}_{\alpha,2}^n(t) \) weakly converge to 0 when \( \alpha > 1/2 \) and to \( \sqrt{W(t)} \), where \( W(t) \) is a standard Brownian motion when \( \alpha = 1/2 \). To see this, we focus on \( \hat{M}_{\alpha,2}^n(t) \). The process \( M_{\alpha,2}^n(t) \) is a compensated Poisson process and is a martingale with predictable quadratic variation

\[
\sum_{k=1}^{N^n_k} \mu_k^n \left( t - \int_0^t \hat{I}_k^n(s) ds \right) = \frac{N^n_{\alpha} \sum_{k=1}^{N^n_k} \mu_k^n t}{n^{2\alpha}} - \frac{\int_0^t \hat{I}_k^n(s) ds}{n^{\alpha}}.
\]

The stochastic boundedness of \( \hat{I}_\alpha^n \) in Lemma 3.3 implies that second term on the right-hand side converges 0 for all \( 1/2 \leq \alpha \leq 1 \). Using the law of large numbers, the first term converges to 0 if
\( \alpha > 1/2 \) and to \( \tilde{\lambda} \) if \( \alpha = 1/2 \). The jumps of \( M_{n,2}^{\alpha}(t) \) are bounded by \( 1/n^{\alpha} \), hence the martingale central limit theorem implies that \( M_{n,2}^{\alpha}(t) \to \sqrt{\lambda} W(t) \). The proof for \( M_{n,3}^{\alpha}(t) \) follows the same steps. Similarly, \( \tilde{M}_{n,3}^{\alpha}(t) \to 0 \) for all \( \alpha \geq 1/2 \) as a result of the stochastic boundedness in Lemma 3.3. Plugging in \( (3.1) \), we get \( n^{-\alpha} \langle \lambda, -N_{n}^{\alpha} \mu F \rangle t \to -\beta \lambda^{\alpha} \mu F^{-\alpha} t \). Finally, using the central limit theorem, we have \( n^{-\alpha} \left( \sum_{k=1}^{n} \tilde{\mu}_{F}^{\alpha} - N_{n}^{\alpha} \mu F \right) t \overset{P}{\to} 0 \) for \( \alpha > 1/2 \) and \( n^{-\alpha} \left( \sum_{k=1}^{n} \tilde{\mu}_{F}^{\alpha} - N_{n}^{\alpha} \mu F \right) t \to \zeta \lambda^{\alpha} \mu F^{-\alpha} t \) for \( \alpha = 1/2 \), where \( \zeta \) is a normal random variable with mean 0 and variance \( \sigma_{\tilde{\mu}}^{2} \). As \( S_{n} \eta_{n,t} \Rightarrow S_{e} \eta_{n,t} \) for any \( \epsilon > 0 \), using Lemma 2 in [BQ22], a modification of the Skorokhod representation theorem, we can assume that all the above processes converge with probability 1 and to prove the theorem, we need to prove for any \( \rho > 0 \), \( \mathbb{P}(d_{S}(\hat{X}_{n}^{\alpha}, \xi_{\alpha}) > \rho) \to 0 \) as \( n \to \infty \), where \( d_{S}(\cdot, \cdot) \) is the metric for Skorokhod-J_{1} topology.

For any \( \varpi > 0 \), we can find a sequence of homeomorphisms \( \Lambda^{n}(t) : [0, T] \to [0, T] \) with derivative \( \dot{\Lambda}^{n}(t) \) and \( N_{\varpi} \) such that for any \( n > N_{\varpi} \),

\[
\begin{align*}
\left| \hat{M}_{n,1}^{\alpha}(t) + \hat{M}_{n,2}^{\alpha}(t) + \hat{M}_{n,3}^{\alpha}(t) - \sqrt{2\lambda} W(\Lambda^{n}(t)) \right|_{T}^{\star} & \geq \left| \langle t, S_{n} \eta_{n,t} \rangle - \langle t, S_{n} \eta_{n,t} \rangle_{T}\right|_{T}^{\star} \geq \left| \hat{\Lambda}^{n}(t) - 1 \right|_{T}^{\star} < \varpi
\end{align*}
\]

if \( \alpha = 1/2 \) and

\[
\begin{align*}
\left| \hat{M}_{n,1}^{\alpha}(t) + \hat{M}_{n,2}^{\alpha}(t) + \hat{M}_{n,3}^{\alpha}(t) \right|_{T}^{\star} & \geq \left| \langle t, S_{n} \eta_{n,t} \rangle - \langle t, S_{n} \eta_{n,t} \rangle_{T}\right|_{T}^{\star} \geq \left| \hat{\Lambda}^{n}(t) - 1 \right|_{T}^{\star} < \varpi
\end{align*}
\]

if \( 1/2 < \alpha \leq 1 \). We will prove our result by showing

\[
\sup_{0 \leq t \leq T} \left| \hat{X}_{n}^{\alpha}(t) - \xi_{\alpha}(\Lambda^{n}(t)) \right|_{T} \to 0, \text{ w.p. 1.} \quad (A.4)
\]

Using the tightness of the processes, without loss of generality, we can assume that

\[
\sup_{n \in \mathbb{N}} \left\{ \left| \langle t, \eta_{n,t} \rangle_{T} \right|_{T}^{\star} \right\} < K
\]

Now, taking \( \xi_{\alpha}(t) \) to be the solution of the appropriate equation in the statement of the theorem, plugging in the definition of fairness process for the seventh term on the right-hand side of (A.3), for any \( \varpi \) we have an \( N_{\varpi} \) such that \( n > N_{\varpi} \) implies

\[
\begin{align*}
|\hat{X}_{n}^{\alpha}(t) - \xi(\Lambda^{n}(t))| & \leq \varpi + \gamma \left| \int_{0}^{t} (\hat{X}_{n}^{\alpha}(s))^{\downarrow} ds - \int_{0}^{\Lambda^{n}(t)} (\xi(s))^{\downarrow} ds \right|
+ \left| \langle t, \eta_{n,t} \rangle \int_{0}^{t} (\hat{X}_{n}^{\alpha}(s))^{\downarrow} ds - \langle t, \eta_{n,t}^{\alpha} \rangle \int_{0}^{\Lambda^{n}(t)} (\xi(s))^{\downarrow} ds \right|
\end{align*}
\]

We can bound the second term on the right-hand side of (A.5) as

\[
\begin{align*}
\left| \int_{0}^{t} (\hat{X}_{n}^{\alpha}(s))^{\downarrow} ds - \int_{0}^{\Lambda^{n}(t)} (\xi(s))^{\downarrow} ds \right| & \leq \int_{0}^{t} \left| \hat{X}_{n}^{\alpha}(s) - \xi_{\alpha}(\Lambda^{n}(s)) \right| ds + \int_{0}^{t} \left| (1 - \Lambda^{n}(t)) \xi_{\alpha}(\Lambda^{n}(s)) \right| ds
\leq \int_{0}^{t} \left| \hat{X}_{n}^{\alpha}(s) - \xi_{\alpha}(\Lambda^{n}(s)) \right| ds + \varpi K t
\end{align*}
\]

24
To bound the third term on the right-hand side of (A.5),
\[
\left|\langle t, \eta_{0,t}^n\rangle \int_0^t (\hat{X}_\alpha^n(s))^- ds - \langle t, \eta_{0,\Lambda^n(t)}\rangle \int_0^\Lambda(t) (\xi(s))^- ds\right|
\]
\[
\leq \left|\left(\langle t, \eta_{0,t}^n\rangle - \langle t, \mathcal{S}_t \eta_{0,t}^n\rangle\right) \int_0^t (\hat{X}_\alpha^n(s))^- ds - \left(\langle t, \eta_{0,\Lambda^n(t)}\rangle - \langle t, \mathcal{S}_t \eta_{0,\Lambda^n(t)}\rangle\right) \int_0^\Lambda(t) (\xi(s))^- ds\right|
\]
\[
+ \left|\left(\langle t, \mathcal{S}_t \eta_{0,t}^n\rangle - \langle t, \eta_{0,\Lambda^n(t)}\rangle\right) \int_0^t (\hat{X}_\alpha^n(s))^- ds - \langle t, \mathcal{S}_t \eta_{0,\Lambda^n(t)}\rangle \int_0^t (\hat{X}_\alpha^n(s))^- ds - \langle t, \mathcal{S}_t \eta_{0,\Lambda^n(t)}\rangle \int_0^t \left(\hat{X}_\alpha^n(s) - (\xi(\Lambda(s)))^-\right) ds\right|
\]
\[
\leq (2\epsilon + \varpi(1 + K)) K t + K \int_0^t \left|\hat{X}_n(s) - \hat{\xi}(\Lambda(s))\right| ds.
\]
Plugging these bounds in (A.5), using Gronwall’s inequality, and choosing \(\epsilon\) and \(\varpi\) appropriately, our result follows. 

Now, we prove the interchangeability of many-server limit and limit as \(t \to \infty\). To do so, we need the following uniform integrability result.

**Lemma A.1.** For any \(1/2 \leq \alpha \leq 1\), the stationary scaled system lengths \(\{\hat{X}_\alpha^n(\infty)\}_{n \in \mathbb{N}}\) are uniformly integrable and hence tight.

**Proof.** For any \(n \in \mathbb{N}\), we decompose the scaled system length process into its negative and positive parts as
\[
\hat{X}_\alpha^n(\infty) = \left(\hat{X}_\alpha^n(\infty)\right)^+ - \left(\hat{X}_\alpha^n(\infty)\right)^-
\]
and prove the uniform integrability of each part separately. We consider \(\left(\hat{X}_\alpha^n(t)\right)^-\) and use a coupling argument. Now, suppose we fix \(n > 0\) and let \(\theta^i_{A_n}\) be the occurrence time of the \(i\)th event for Poisson process \(A^n(t)\). Assume that we know \((\tilde{\mu}_1, \ldots, \tilde{\mu}_n)\) and define \(\hat{S}^n(t)\) to be a Poisson process with rate \(\sum_{k=1}^n \tilde{\mu}_k\) and let \(\theta^i_{S^n,*}\) be the occurrence time for the \(i\)th event. We also define a sequence \(\{U^n_{i}\}_{i \in \mathbb{N}}\) of independent uniform(0,1) random variables. For any given time \(t\), we know that the system should be serving with a rate equal to the sum of service rates of busy servers, i.e., \(\sum_{k=1}^n \tilde{\mu}_k(1 - I^n_k(t))\), hence we use the thinning property of Poisson process where event \(i\) of \(\hat{S}^n(t)\) is accepted as an actual departure with probability \(p^n_i(\theta^i_{S^n,*}) = \frac{\sum_{k=1}^n \tilde{\mu}_k(1 - I^n_k(\theta^i_{S^n,*}))}{\sum_{k=1}^n \tilde{\mu}_k}\) by checking whether \(I^n_k \leq p^n_i(\theta^i_{S^n,*})\). The processes \(I^n_k(t)\) can be rigorously defined by using \(U^n_{i}\)’s and a routing process based on our routing policy. As this does not play a major role in our proof, we refer the reader to [BQ22] for the detailed construction of idleness processes. Now, we can write
\[
\left(\hat{X}_\alpha^n(t)^-\right)^- = \left(\hat{X}_\alpha^n(0)^-\right)^- + n^{-\alpha} \sum_{i=1}^{\hat{S}^n(t)} \mathbb{I}\left(\left(\hat{X}_\alpha^n(\theta^i_{S^n,*})^+ = 0\right) \land \left(U^n_i \leq p^n_i(\theta^i_{S^n,*})\right)\right) - n^{-\alpha} \sum_{i=1}^{\Lambda^n(t)} \mathbb{I}\left(\left(\hat{X}_\alpha^n(\theta^i_{A^n,*})^-\right)^- > 0\right).
\]
We define a birth-death process \( \{Y^n_i(t)\}_{n \in \mathbb{N}} \) with \( Y^n_i(0) = (X^n(0))^- \) w.p.1, whose birth rate at \( Y^n_i(t) = i \) is \( \sum_{k=1}^{\infty} \hat{\mu}_k - \mu_{\min} \) and death rate is \( \lambda^n \) if \( Y^n_i(t) > 0 \). Then, we can couple the scaled process \( \tilde{Y}^n_i(t) = n^{-\alpha} Y^n_i(t) \) with the system length process by writing it as

\[
\tilde{Y}^n_i(t) = \tilde{Y}^n_i(0) + n^{-\alpha} \sum_{i=1}^{n\alpha} \mathbb{1}(U^n_i \leq \tilde{p}^n_i(\theta^n_{\alpha i} \leq -)) - n^{-\alpha} \sum_{i=1}^{A^n(t)} \mathbb{1}(\tilde{Y}^n_i(\theta^n_{A^n(t)} > 0)),
\]

where \( \tilde{p}^n_i(\theta^n_{\alpha i} \leq -) := \frac{\sum_{k=1}^{n\alpha} \hat{\mu}_k - \mu_{\min} \hat{Y}_i(\theta^n_{\alpha i} \leq -)}{\sum_{k=1}^{n\alpha} \hat{\mu}_k} \). To see that \( \tilde{X}^n_i(t) \leq \tilde{Y}^n_i(t) \) for all \( t \geq 0 \) with probability 1, define \( \vartheta^n = \left\{ t : \left( \tilde{X}^n_i(t) \right)^- \leq \tilde{Y}^n_i(t) \right\} \). As at most one event occurs at any given time \( t \) with probability 1, we have \( \left( \tilde{X}^n_i(\vartheta^n \leq -) \right) ^- = \tilde{Y}^n_i(\vartheta^n \leq -) \). By definition we have

\[
\tilde{p}^n_i(\vartheta^n \leq -) \geq p^n_i(\vartheta^n \leq -).
\]

Hence, if \( \vartheta^n \) is an event epoch for \( \tilde{S}^n(t) \), we have \( \left( \tilde{X}^n_i(\vartheta^n) \right) ^- \leq \tilde{Y}^n_i(\vartheta^n) \) and if \( \vartheta^n \) is an event epoch for \( A^n(t) \), we have \( \left( \tilde{X}^n_i(\vartheta^n) \right) ^- = \tilde{Y}^n_i(\vartheta^n) \), both contradicts with the definition of \( \vartheta^n \). Hence, we conclude that \( \left( \tilde{X}^n_i(\vartheta^n \leq -) \right) ^- \leq \tilde{Y}^n_i(t) \) for all \( t \geq 0 \) with probability 1.

Now, define \( \zeta^n := n^{-\alpha} \left( \sum_{k=1}^{n\alpha} \hat{\mu}_k - N^n \tilde{\mu} \right) \) and by re-arranging the terms we have

\[
\sum_{k=1}^{n\alpha} \hat{\mu}_k = \lambda^n + \beta (\lambda^n)^{-1} (\hat{\mu}^{-1} + n^{-\alpha} \zeta^n).
\]

Take \( M_1 := (\beta (\lambda^n)^{-1} \hat{\mu}^{-1} + n^{-\alpha} \zeta^n)^{+} + 2n^{-\alpha} \), and define a new birth-death process \( Y^n_q(0) = Y^n_i(0) \) with probability 1, whose birth rate at \( Y^n_q(0) = i \) is \( \lambda^n - \mu_{\min} \min\{i, M_1\} \) and death rate is \( \lambda^n \) at \( Y^n_q(0) \neq 0 \). By definition \( Y^n_q(t) \) is stochastically greater than \( Y^n_i(t) \). Using a similar argument as above, we can couple \( Y^n_q(t) \) with a simple birth death process \( Y^n_m(t) \) where \( Y^n_m(0) = \left( Y^n_q(0) - M \right)^{+} \) with birth rate \( \lambda^n - \mu_{\min} M_1 \), death rate \( \lambda^n \) and \( Y^n_m(t) \leq M_1 + Y^n_m(t) \) for all \( t \geq 0 \). As birth and death rates are constants, \( Y^n_m(t) \) is equivalent to an \( M/M/1 \) queue. Hence, we can prove that \( \left( \tilde{X}^n(\infty) \right) ^- \) is uniformly integrable by showing that \( \sup \mathbb{E} \left[ (n^{-\alpha}(M_1 + Y^n_m(\infty)))^2 \right] < \infty \). For \( n \in \mathbb{N} \),

\[
\mathbb{E} \left[ (n^{-\alpha}(M_1 + Y^n_m(\infty)))^2 \right] = n^{-2\alpha} \left( \mathbb{E}[M^2_1] + 2\mathbb{E}[M_1 Y^n_m(\infty)] + \mathbb{E}[Y^n_m(\infty)]^2 \right)
\]

\[
= n^{-2\alpha} \left( \mathbb{E}[M^2_1] + 2\mathbb{E}[M_1 \mathbb{E}[Y^n_m(\infty)] + \mathbb{E}[Y^n_m(\infty)]^2 \mathbb{E}[\tilde{\mu}^2]\right)
\]

\[
= n^{-2\alpha} \left( \mathbb{E}[M^2_1] + 2 \mathbb{E} \left[ \frac{\lambda^n - \mu_{\min} M_1}{\mu_{\min}} \right] + \mathbb{E} \left[ \frac{(\lambda^n - \mu_{\min} M_1)(2\lambda^n - \mu_{\min} M_1)}{\mu_{\min}^2 M_1^2} \right] \right)
\]

\[
= n^{-2\alpha} \left( \mathbb{E}[M^2_1] + 2 \frac{\lambda^n}{\mu_{\min}} - 2\mathbb{E}[M_1] + 2 \frac{(\lambda^n)^2}{\mu_{\min}^2 M_1^2} - 3\mathbb{E} \left[ \frac{\lambda^n}{\mu_{\min} M_1} + 1 \right] \right)
\]

\[
\leq n^{-2\alpha} \left( \mathbb{E}[M^2_1] + 2 \frac{\lambda^n}{\mu_{\min}} + 2\mathbb{E}[M_1] + \frac{(\lambda^n)^2}{\mu_{\min}^2 M_1^2} + 3\lambda^n \frac{\mu_{\min} M_1}{\mu_{\min}^2 M_1^2} + 1 \right)
\]

Assumption 3.1 implies that the second, fourth and fifth terms converges to a finite number for any \( \alpha \geq 1/2 \) and hence can be bounded uniformly for any \( n \). Also, using the simple identity \((a + b)^2 \leq \)
\[ 2a^2 + 2b^2, \text{ we have} \]
\[ n^{-2\alpha}E[M_1^2] \leq 4n^{-2\alpha} \beta^2 (\lambda^n)^{2\alpha} \tilde{\mu}^{2-2\alpha} + 4E[(\zeta^n)^2] + 2. \]

Again we can use Assumption 3.1 to show that the first term converges to a finite number. Using independence
\[ E[(\zeta^n)^2] = n^{-2\alpha}E \left[ \sum_{k=1}^{N^n} (\hat{\mu}_k^n - N^n \tilde{\mu})^2 \right] = n^{-2\alpha}E \left[ \sum_{k=1}^{N^n} (\hat{\mu}_k^n - \bar{\mu})^2 \right] \leq n^{-2\alpha}N^n (\mu_{\max} - \mu_{\min})^2, \]

which also converges and hence can be uniformly bounded for all \( n \). Similarly,
\[ n^{-2\alpha}E[M_1] \leq n^{-2\alpha}E[\beta (\lambda^n)^{\alpha} \tilde{\mu}^{1-\alpha} |\mu_{\max} - \mu_{\min}|] + 2n^{-\alpha} - n^{-2\alpha}N^n, \]

which also converges and proves the uniform integrability of \( \{X^n(\infty)\} \). The proof of the uniform integrability of \( \{\hat{X}_n(\infty)\} \) follows the same lines and hence is omitted. \( \square \)

**Proof of Theorem 4.2** Suppose \( X^n(0), (\tilde{\mu}_1^n, \ldots, \tilde{\mu}_{N^n}) \) are distributed according to the stationary measure of the \( n \)th system \( \pi^n \). Then, \( X^n(t), (\tilde{\mu}_1^n, \ldots, \tilde{\mu}_{N^n}) \) are also distributed according to \( \pi^n \). As we know that the \( \{\pi^n\} \) are tight and \( \hat{X}_n(\infty) = \xi_n(t) \), the first convergence holds and the second convergence holds as a result of the uniform integrability. To see the third one, we need to see
\[ \lim_{T \to \infty} \frac{1}{T} \sum_{k=1}^{N^n} \delta_{\tilde{\mu}_k^n} (\tilde{\theta}_k^n) \int_0^T I_k^n(s) ds = \sum_{k=1}^{N^n} \delta_{\tilde{\mu}_k^n} (\tilde{\theta}_k^n) E[I_k^n(\infty) | \bar{\mu}]. \]

Again, starting with a stationary system and using the uniform integrability the result follows. \( \square \)

**A.4 Proofs for Results Presented in Section 5**

**Proof of Lemma 5.3** We define the discrete-time measure-valued stochastic process
\[ \mathcal{U}^n_{A,i}(\tilde{\theta}) = \delta_{\mu_k}(\tilde{\theta}) \mathbb{I}(X^n(\theta_{A,i} - ) \leq 0) \text{ for all } \tilde{\theta} \in \mathcal{B}(\mathbb{R}_+), \]

if the \( i \)th incoming arrival in the \( n \)th system is immediately routed to server \( k \) and 0 (thought as a measure) otherwise. Similarly, we define \( \mathcal{U}^n_{S,i}(\tilde{\theta}) = \delta_{\mu_k}(\tilde{\theta}) \mathbb{I}(X^n(\theta_{S,i} - ) \leq 0) \) for all \( \tilde{\theta} \in \mathcal{B}(\mathbb{R}_+) \), if the \( i \)th potential service completion is from \( k \)th server, if it is not an actual service completion \( \mathcal{U}^n_{S,i} = 0 \). Then, for any \( f \in C^0 \) we have the following balance
\[ \langle f, \tilde{\psi}^{n} \rangle = \langle f, \tilde{\psi}^{0} \rangle + n^{-1} \sum_{i=1}^{S^n(t)} \langle f, \mathcal{U}^n_{A,i} \rangle - n^{-1} \sum_{i=1}^{A^n(t)} \langle f, \mathcal{U}^n_{S,i} \rangle. \]  

To prove tightness, we use the conditions introduced by [Jak86].

**Theorem A.2** ([Jak86]). A sequence of stochastic processes \( \{\eta^n_t\} \) taking values in \( \mathbb{D}_P[0,T] \) is tight if and only if:
J1. (Compact Containment Condition) For each \( \rho, T > 0 \), there exists a compact set \( K_\rho \subset P \) such that
\[
\liminf_{n \to \infty} P(\tilde{\psi}^n_t \in K_\rho, \text{ for all } t \in [0, T]) > 1 - \rho.
\]

J2. There exists a family of functions \( F \) such that
\[ i. \quad H \in F : P \to \mathbb{R}, F \text{ separates points in } P \text{ and } F \text{ is closed under addition.} \]
\[ ii. \quad \text{For any fixed } H \in F, \text{ the sequence of functions } \{h^n(t) := H(\eta^n_t), \text{ for all } t \in [0, T]\} \text{ is tight in } D_{\mathbb{R}}[0, \infty) \text{ endowed with Skorokhod-J_1 topology.} \]

Using Lemma 3.3, we know that, for all \( \epsilon > 0 \), there exists a \( K_\epsilon \) such that \( P(\sup_{0 \leq t \leq T} |\hat{I}_n(t)| > K_\epsilon) < \epsilon \) for all \( n \in \mathbb{N} \). Define \( K_\epsilon \) as the set of measures bounded by \( K_\epsilon \) on the support \([\mu_{\min}, \mu_{\max}]\).

The set \( K_\epsilon \) is compact and Lemma 3.3 implies J1.

To show that J2 holds, we define
\[
F = \{ H : M_F[\mu_{\min}, \mu_{\max}] \to \mathbb{R} : \exists f \in C^b_{\mathbb{R}^+}[0, \infty) \text{ such that } H(\psi) = \langle f, \psi \rangle \text{ for all } \psi \in M_F \},
\]
where \( M_F[\mu_{\min}, \mu_{\max}] \) is the set of finite measures on \([\mu_{\min}, \mu_{\max}]\). The set \( F \) separates the points in \( M_F \) and is closed under addition. Take \( K_f \) such that \( H(\psi) = \langle f, \psi \rangle \) with \( f(\mu) \leq K_f \) for all \( \mu \in [\mu_{\min}, \mu_{\max}] \). To show tightness of \( \{(f, \tilde{\psi}^n_t)\} \), we need to show that for all \( \epsilon, \rho > 0 \)

1. there exists an \( M_{f,\epsilon} \) such that \( P(\sup_{0 \leq t \leq T} |\langle f, \tilde{\psi}^n_t \rangle| > M_{f,\epsilon}) < \epsilon \) and

2. there exists a \( \rho \) and an \( N_\rho \) such that for all \( n > N_\rho \), \( P(\omega(f, \tilde{\psi}^n), \rho \geq \epsilon) < \epsilon \), where
\[
w(f, \tilde{\psi}^n_i, \rho) = \inf \max_{\{t_i\}} \sup_{t_i \leq t \leq t_{i+1}} |\langle f, \tilde{\psi}^n_t \rangle - \langle f, \tilde{\psi}^n_s \rangle|
\]
and \( \{t_i\}_{0 \leq i \leq \nu} \) is any \( \rho \)-sparse set, i.e., \( 0 = t_0 < t_1 < \cdots < t_\nu = T \) with \( \min_i |t_{i+1} - t_i| > \rho \).

Again taking \( M_{f,\rho} = K_fK_f \), we have
\[
P\left(\sup_{0 \leq t \leq T} |\langle f, \tilde{\psi}^n_t \rangle| > M_{f,\rho}\right) \leq P\left(K_f \sup_{0 \leq t \leq T} \hat{I}_n(t) > M_{f,\rho}\right) \leq \rho,
\]
which implies the first condition. Now, we prove the second condition. Using (A.6), for any \( 0 \leq s < \nu \),
t \leq T \text{ we have} \]

\[
|\langle f, \bar{\psi}_t^n \rangle - \langle f, \bar{\psi}_s^n \rangle| = n^{-1} \sum_{i=1}^{S^n(t)} \langle f, U_{S,i}^n \rangle - n^{-1} \sum_{i=1}^{A^n(t)} \langle f, U_{A,i}^n \rangle - n^{-1} \sum_{i=1}^{S^n(s)} \langle f, U_{S,i}^n \rangle - n^{-1} \sum_{i=1}^{A^n(s)} \langle f, U_{A,i}^n \rangle \\
\leq n^{-1}K_f \left| S^n(t) - S^n(s) - \sum_{k=1}^{N^n} \mu_k t + \sum_{k=1}^{N^n} \mu_k s \right| + n^{-1}K_f \sum_{k=1}^{N^n} \mu_k |t - s| \\
+ n^{-1}K_f |A^n(t) - A^n(s) - \lambda^n t + \lambda^n s| + n^{-1}K_f \lambda^n |t - s| \\
\leq n^{-1}K_f \left| S^n(t) - \sum_{k=1}^{N^n} \mu_k t \right| + n^{-1}K_f \left| S^n(s) - \sum_{k=1}^{N^n} \mu_k s \right| \\
+ K_f n^{-1} \sum_{k=1}^{N^n} \mu_k - \lambda(1 + \beta) \left| t - s \right| + n^{-1}K_f \lambda^n |t - s| \\
+ n^{-1}K_f \lambda^n |t - s| + \bar{\lambda}(1 + \beta) |t - s| \\
\leq n^{-1}K_f \sup_{0 \leq t \leq T} \left| S^n(t) - \sum_{k=1}^{N^n} \mu_k t \right| + n^{-1}K_f \sup_{0 \leq t \leq T} |A^n(t) - \lambda^n t| \\
+ K_f n^{-1} \sum_{k=1}^{N^n} \mu_k - \lambda(1 + \beta) \left| t - s \right| + n^{-1}K_f \lambda^n |t - s| + \bar{\lambda}(1 + \beta) |t - s| \\
(A.7)
\]

Using the martingale central limit theorem, the first and second terms on the right-hand side converge to 0 in probability. Similarly, using the law of large numbers, we can show that the third term also converges to 0 in probability. Finally, from Assumption 3.1 we know that $n^{-1} \lambda^n \rightarrow \bar{\lambda}$. Hence, choosing $\rho < \epsilon/2(K_f \bar{\lambda}(2 + \beta))$ and $N$ large enough the second condition follows. Moreover, by examining (A.7), one can conclude that any limit is continuous. 

\begin{proof}[Proof of Lemma 5.4] The martingale central limit theorem (c.f. [PTW07]) implies $\sup_{0 \leq t \leq T} n^{-1} |S^n(t) - \sum_{k=1}^{N^n} \mu_k t| \xrightarrow{P} 0$ and $\sup_{0 \leq t \leq T} n^{-1} |A^n(t) - \lambda^n t| \xrightarrow{P} 0$ Using the Skorokhod representation theorem (see, e.g., Theorem 6.7 in [Bil99]), we can assume that these and the subsequent in the statement of the lemma converge almost surely. Also, as mentioned in the proof of Lemma 5.3 for any $f \in C_{\mathbb{R}^+}[0, T]$, the limit $\langle f, \bar{\psi}_t \rangle$ is continuous and hence, the convergence holds in the supremum norm as well as the Skorokhod $d_S$ metric. We define the processes

\[
M^n_1(t) := \sum_{i=1}^{S^n(t)} \langle f, U_{S,i}^n \rangle - \sum_{i=1}^{A^n(t)} \langle f, U_{A,i}^n \rangle - \sum_{k=1}^{N^n} \mu_k \left( f \cdot \frac{1 - I_k^n(\theta_{S,i}^n - \cdot)}{\sum_{k=1}^{N^n} \mu_k} \mathbb{I}(X^n(\theta_{S,i}^n - \cdot) \leq 0) \right) \\
M^n_2(t) := \sum_{i=1}^{A^n(t)} \langle f, U_{A,i}^n \rangle - \sum_{i=1}^{A^n(t)} \langle f, U_{A,i}^n \rangle - \sum_{k=1}^{N^n} \mu_k \left( f \cdot \frac{1 - I_k^n(\theta_{A,i}^n - \cdot)}{\sum_{k=1}^{N^n} \mu_k} \mathbb{I}(X^n(\theta_{A,i}^n - \cdot) \leq 0) \right),
\]

where 0/0 is assumed to be 0. It is easy to see that both $M^n_1$ and $M^n_2$ are $\mathcal{F}_t$ martingales. After some
algebraic manipulations, equation (A.6) becomes

\[
\langle f, \tilde{v}_t^n \rangle = \langle f, \tilde{v}_0^n \rangle + n^{-1} M_1^n(t) + n^{-1} \sum_{i=1}^{S^n(t)} \left\langle f, \frac{\sum_{k=1}^{N^n} \mu_k (1 - I^n_k (\theta^n_{S,i} - )) \delta_{\mu_k} I(X^n(\theta^n_{S,i} - ))}{\sum_{k=1}^{N^n} \mu_k} \right\rangle \\
- n^{-1} M_2^n(t) - n^{-1} \sum_{i=1}^{A^n(t)} \left\langle f, \frac{\sum_{k=1}^{N^n} h(\mu_k) I^n_k (\theta^n_{A,i} - ) \delta_{\mu_k} I(X^n(\theta^n_{A,i} - ))}{\sum_{k=1}^{N^n} h(\mu_k) I^n_k (\theta^n_{S,i} - )} \right\rangle \\
= \langle f, \tilde{v}_0^n \rangle + n^{-1} M_1^n(t) + n^{-1} \left( \langle f, t, \sum_{k=1}^{N^n} \delta_{\mu_k} \right\rangle \int_0^t I(X^n(s - ) \leq 0) dS^n(s) \\
- n^{-1} \int_0^t \left\langle f, t, \tilde{v}_s^n \right\rangle I(X^n(s - ) \leq 0) dS^n(s) - n^{-1} M_2^n(t) \\
- n^{-1} \left( \int_0^t \frac{\langle f, t, \tilde{v}_s^n \rangle - \langle f, \tilde{v}_t^n \rangle}{(h, \tilde{v}_s^n)} \right) dA^n(s) \\
= \langle f, \tilde{v}_0^n \rangle + \frac{M_1^n(t)}{n} + \frac{\langle f, t, \sum_{k=1}^{N^n} \delta_{\mu_k} \right\rangle \int_0^t I(X^n(s - ) \leq 0)\left( \frac{X^n(s - )}{n} \right)}{n} \int_0^t I(X^n(s - ) \leq 0) d\left( \frac{S^n(s) - \sum_{k=1}^{N^n} \mu_k s}{n} \right) \\
- \int_0^t \left( \frac{\langle f, t, \tilde{v}_s^n \rangle}{(h, \tilde{v}_s^n)} \right) I(X^n(s - ) \leq 0) d\left( \frac{S^n(s) - \sum_{k=1}^{N^n} \mu_k s}{n} \right) \\
- \int_0^t \left( \frac{\langle f, t, \tilde{v}_s^n \rangle}{(h, \tilde{v}_s^n)} \right) I(X^n(s - ) \leq 0) d\left( \frac{A^n(s) - \lambda^n s}{n} \right) \\
- \int_0^t \left( \frac{\langle f, t, \tilde{v}_s^n \rangle}{(h, \tilde{v}_s^n)} \right) I(X^n(s - ) \leq 0) d\left( \frac{\lambda^n s}{n} \right) (A.8)
\]

Since \( f \in C^b_{[\mu_{\min}, \mu_{\max}]}[0, \infty) \), we assume that \( f(\mu) \leq K_f \) for all \( \mu \in [\mu_{\min}, \mu_{\max}] \). By our assumption, we know that \( \sup_{0 \leq t \leq T} |\langle f, \tilde{v}_t^n, k \rangle - \langle f, \tilde{v}_t \rangle| \to 0 \) almost surely, along the subsequence \( \{\tilde{v}_t^n\}_{k=1}^\infty \). The martingales \( M_1^n(t) \) and \( M_2^n(t) \) can be written as

\[
n^{-1} M_1^n(t) = n^{-1} \sum_{i=1}^{S^n(t)} \langle f, U_{S,i}^n \rangle - \mathbb{E}[\langle f, U_{S,i}^n \rangle | \mathcal{F}_{t^-}] \\
n^{-1} M_2^n(t) = n^{-1} \sum_{i=1}^{A^n(t)} \langle f, U_{A,i}^n \rangle - \mathbb{E}[\langle f, U_{A,i}^n \rangle | \mathcal{F}_{t^-}].
\]

Since both are pure jump martingales, we can write the optional quadratic variation of these mar-
Hence, we have the following bounds:

\[
[n^{-1}M_1^n(t)] = n^{-2} \sum_{i=1}^{S^n(t)} \left( (f, U_{C,i}^n) - \mathbb{E}[(f, U_{C,i}^n) | F_{C,i}^n] \right)^2 \leq \frac{K^2}{n^2} \sup_{0 \leq t \leq T} S^n(t)
\]

\[
[n^{-1}M_2^n(t)] = n^{-2} \sum_{i=1}^{A^n(t)} \left( (f, U_{A,i}^n) - \mathbb{E}[(f, U_{A,i}^n) | F_{A,i}^n] \right)^2 \leq \frac{K^2}{n^2} \sup_{0 \leq t \leq T} A^n(t)
\]

which converges to 0 almost surely. We know that \( h(\mu) > 0 \) and continuous on the closed interval \([\mu_{\min}, \mu_{\max}]\) and hence, there exists \( \epsilon_h \) and \( K_h \) such that \( 0 < \epsilon_h \leq h(\mu) \leq K_h \) for all \( \mu \in [\mu_{\min}, \mu_{\max}] \).

Also, Assumption 3.1 implies the existence of \( K_N \) and \( k_n \) such that \( 0 < \epsilon_N \leq n^{-1}N^n \leq K_N < \infty \). Hence, we have the following bounds:

\[
\begin{align*}
\langle f \times t, \bar{\psi}^n \rangle &\leq \langle f \times t, n^{-1} \sum_{k=1}^{N^n} \delta_{\mu_k} \rangle \leq K_f K_N \mu_{\max}, \\
\langle f \times h, \bar{\psi}^n \rangle &\leq \langle f \times h, n^{-1} \sum_{k=1}^{N^n} \delta_{\mu_k} \rangle \leq K_f K_h K_N
\end{align*}
\]

\[
\left( \frac{f \times \iota, \bar{\psi}^n}{n^{-1} \sum_{k=1}^{N^n} \mu_k} \right) \leq \frac{K_f K_N \mu_{\max}}{\epsilon_N} \quad \text{and} \quad \left( \frac{f \times h, \bar{\psi}^n}{\langle h, \bar{\psi}^n \rangle} \right) \leq \frac{K_f K_h}{\epsilon_h}.
\]

These bounds along with \( \sup_{0 \leq t \leq T} n^{-1} |S^n(t) - \sum_{k=1}^{N^n} \mu_k t| \to 0 \) and \( \sup_{0 \leq t \leq T} n^{-1} |A^n(t) - \lambda^n t| \to 0 \)
almost surely as \( n \to \infty \), the third, fifth and the eighth terms on the right-hand side of (A.8) converges to 0 almost surely. Using the dominated convergence theorem, we deduce that

\[
\begin{align*}
\int_0^t \langle f \times \iota, \bar{\psi}^n \rangle \frac{X^n(s-)}{n} \leq 0 \rangle ds &\to \int_0^t \langle f \times \iota, \bar{\psi} \rangle \mathbb{I}(\xi(s-) \leq 0) ds \\
\int_0^t \langle f \times h, \bar{\psi}^n \rangle \frac{X^n(s-)}{n} \leq 0 \rangle ds &\to \int_0^t \langle f \times h, \bar{\psi} \rangle \frac{\lambda^n}{\langle h, \bar{\psi} \rangle} \mathbb{I}(\xi(s-) \leq 0) ds.
\end{align*}
\]

Hence, our lemma follows.

**Proof of Theorem 5.5**

We know that \( \xi_{1,\infty} > 0 \), and hence any fixed point of (5.4) satisfies

\[
\bar{\lambda} \frac{(1 + \beta)}{\mu} \langle f \times \iota, F \rangle \mathbb{I}(\xi_{1,\iota} \leq 0) = \langle f \times \iota, \bar{\psi}_{1,\iota} \rangle \mathbb{I}(\xi_{1,\iota} \leq 0) + \bar{\lambda} \frac{\langle f \times h, \bar{\psi}_{1,\iota} \rangle}{\langle h, \bar{\psi}_{1,\iota} \rangle} \mathbb{I}(\xi_{1,\iota} \leq 0).
\]

As (5.2) implies that \( \bar{\psi}_{1,\infty} \) is absolutely continuous, it possesses a Radon-Nikodym derivative \( \bar{g}(\mu) \) with respect to \( F \). Setting \( f(\mu) = 1 \) for all \( \mu \), we have

\[
\int_{\mu_{\min}}^{\mu_{\max}} \mu \bar{g}(\mu) dF(\mu) = \bar{\lambda} \beta.
\]

As the equation holds for any \( f \in C_b[\mu_{\min}, \mu_{\max}] \), defining \( c_g = \int_{\mu_{\min}}^{\mu_{\max}} h(\mu) \bar{g}(\mu) dF(\mu) \) to simplify the notation, for \( F \)-almost all \( \mu \)

\[
\mu \bar{g}(\mu) + \frac{\bar{\lambda} h(\mu) \bar{g}(\mu)}{c_g} = \frac{\bar{\lambda}}{\mu_F}(1 + \beta) \mu.
\]

Defining \( L_F = \bar{\lambda}/c_g \) and re-organizing terms, we obtain

\[
\bar{g}(\mu) = \frac{\bar{\lambda}}{\mu_F}(1 + \beta)(1 + L_F h(\mu))^{-1}.
\]
Plugging in (A.10), we obtain (5.6), and since the integrand is decreasing in $L_F$ then the $L_F$ that satisfies this equation must be unique. This proves that for any limit point $\psi_{1,t}$, $(f, \psi_{1,t}) \rightarrow \int_{\mu_{\text{min}}}^{\mu_{\text{max}}} f(\mu) g(\mu) dF(\mu)$ as $t \rightarrow \infty$ for all $f \in \mathbb{C}^b_{[\mu_{\text{min}}, \mu_{\text{max}}]} [0, \infty)$. Hence, using the fact that indicator function of any Borel set $\mathcal{A}$ can be approximated by functions in $\mathbb{C}^b_{[\mu_{\text{min}}, \mu_{\text{max}}]} [0, \infty)$, we have $\psi_{1,t}(\mathcal{A}) \rightarrow \int_{\mathcal{A}} g(\mu) dF(\mu)$. Using Theorem 4.2 this implies that $\psi_{1,t} (\mu_{\text{min}}) = \int_{\mu_{\text{min}}}^{\mu_{\text{max}}} g(\mu) dF(\mu)$ as $n \rightarrow \infty$. Plugging $\psi_{1,t}$ into the definition of the fairness process, the result follows.

Proof of Lemma 5.6 Dividing both sides of (A.10) by $\lambda_\beta$, this equation defines a probability measure $\tilde{F}$ on $[\mu_{\text{min}}, \mu_{\text{max}}]$ and $c_g$ in the proof of Theorem 5.5 can be written as $c_g = \lambda_\beta \int_{\mu_{\text{min}}}^{\mu_{\text{max}}} h(\mu) \frac{\partial g(\mu)}{\lambda_\beta} dF(\mu)$. The integral is the expectation of $h(\mu)$ where $\mu$ follows $\tilde{F}$, and hence, is bounded below by $h_{\text{min}}$ and above by $h_{\text{max}}$. The result follows as $L_F = \tilde{\lambda}/c_g$.

Proof of Corollary 5.7 Let $\tilde{g}(\mu)$ be the density function of $\tilde{\psi}_{1,\infty}$ with respect to $F$. The definition of the limiting fairness measure and (4.3) implies that $\tilde{g}(\mu) = \beta \lambda (\bar{c}_{\eta_1,\infty})^{-1} g(\mu)$. Plugging into the definition of $c_g$, we can see that for the proposed $h(\mu)$,

$$c_g = \int_{\mu_{\text{min}}}^{\mu_{\text{max}}} \left(1 + \frac{\beta}{\mu_{\text{F}}} - \beta (\bar{c}_{\eta_1,\infty})^{-1} g(\mu) \right) \mu dF(\mu)$$

$$= \frac{1 + \beta}{\mu_{\text{F}}} \int_{\mu_{\text{min}}}^{\mu_{\text{max}}} \mu dF(\mu) - \beta (\bar{c}_{\eta_1,\infty})^{-1} \int_{\mu_{\text{min}}}^{\mu_{\text{max}}} \mu g(\mu) dF(\mu)$$

$$= 1.$$

Hence, $L_F = \tilde{\lambda}$. Now, plugging this in (A.11), we get the desired result.

A.5 Proofs for Results Presented in Section 6

Proof of Theorem 5.2 To simplify notation, define $C = \beta \lambda^{\alpha-1} \bar{\mu}_{\text{F}}^{-\alpha-\alpha}(\bar{c}_{\eta_1})^{-1}$. The uniform convergence assumption implies that for any $\rho > 0$, there exists an $N_{\rho}$ such that for all $\mu_{\text{min}} \leq \mu \leq \mu_{\text{max}}$

$$|\mathbb{E}[\bar{\mu}_{n,k}^n - \mu_{\text{min}}] | \leq n^{\alpha-1} C g(\mu) \leq \alpha^{\alpha-1} \rho \quad (A.12)$$

First, we analyze the best response for server $k$ for all large $n$ for the approximating problem

$$\max_{\mu_{\text{min}}, \mu_{\text{max}}} f(n^{\alpha-1} C g(\mu)) - a_k^n c(\mu). \quad (A.13)$$

Taking the derivative of the objective, we get

$$n^{\alpha-1} g'(\mu) f'(n^{\alpha-1} C g(\mu)) - a_k^n c'(\mu). \quad (A.14)$$

First, we analyze part 1, i.e., the situation when $n^{\alpha-1} f'(n^{\alpha-1} x) \rightarrow 0$ for all $x > 0$. We know that $f(\mu)$ is concave and hence, the derivative is decreasing. Since $g(\mu)$ is continuous, it attains its minimum at $\mu_* \in [\mu_{\text{min}}, \mu_{\text{max}}]$. Then, we have $f'(n^{\alpha-1} C g(\mu_*^n)) \geq f'(n^{\alpha-1} C g(\mu))$ for all $\mu_{\text{min}} \leq \mu \leq \mu_{\text{max}}$. Hence, there exists an $N_1$ such that $n^{\alpha-1} f'(n^{\alpha-1} C g(\mu_*^n)) < \frac{\min_{\mu_{\text{min}}} c(\mu) - \min_{\mu_{\text{max}}} c(\mu)}{2C}$ for all $n > N_1$ and hence, the minimizer of (A.13) is $\mu_*^n = \mu_{\text{min},k}$ for all $k$ and $n > N_1$. Using the gradient inequality for concave functions on $f(n^{\alpha-1} x)$, we get for any $\mu_{\text{min}} \leq \mu \leq \mu_{\text{max}},$

$$f(n^{\alpha-1} C g(\mu)) - a_k^n c(\mu) \leq f(n^{\alpha-1} C g(\mu^*_{\text{min},k}^n)) + n^{\alpha-1} C f'(n^{\alpha-1} C g(\mu_*^n))(g(\mu) - g(\mu_{\text{min},k}^n))$$

$$- a_k^n c(\mu^*_{\text{min},k}^n) - a_{\min} c_{\min}(\mu - \mu_{\text{min},k}^n).$$
Hence, for any $n > N_1$, we have
\[
 f(n^\alpha C g(\mu)) - \tilde{a}_k^n c(\mu) \leq f(n^\alpha C g(\mu_{\min,k}^n)) - \tilde{a}_k^n c(\mu_{\min,k}^n) - \frac{a_{\min} c_{\min}}{2}(\mu - \mu_{\min,k}^n)
\]
and for any $\mu > \mu_{\min,k}^n + 6\epsilon(a_{\min} c_{\min})^{-1}$,
\[
 f(n^\alpha C g(\mu)) - \tilde{a}_k^n c(\mu) \leq f(n^\alpha C g(\mu_{\min,k}^n)) - \tilde{a}_k^n c(\mu_{\min,k}^n) - 3\epsilon. \tag{A.15}
\]
Now, again using (A.12), there exists an $N_2$ such that $n > N_2$ implies
\[
 \mathbb{E}[I^n_k(\infty)|\tilde{\mu}_k^n] = \mu \geq n^{\alpha-1}C g(\mu^*_g) \frac{2}{\rho}.
\]
Using the mean-value theorem and the concavity of $f(\cdot)$, for any $\mu_{\min} \leq \mu \leq \mu_{\max}$ and $n > N_2 \lor N_\rho$
\[
 |f(\mathbb{E}[I^n_k(\infty)|\tilde{\mu}_k^n] - f(n^{\alpha-1}C g(\mu))| \leq n^{\alpha-1} f'(n^{\alpha-1}C g(\mu^*_g)) \frac{2}{\rho} \rho.
\]
Now, choosing $N_3$ such that $n > N_3$ implies
\[
 n^{\alpha-1} f'(n^{\alpha-1}C g(\mu^*_g)) \frac{2}{\rho} \leq \frac{\epsilon}{\rho},
\]
and using (A.15), we have
\[
 f(\mathbb{E}[I^n_k(\infty)|\tilde{\mu}_k^n] = \mu)) - \tilde{a}_k^n c(\mu) \leq f(\mathbb{E}[I^n_k(\infty)|\tilde{\mu}_k^n = \mu]) - \tilde{a}_k^n c(\mu_{\min,k}^n) - \epsilon,
\]
for all $\mu > \mu_{\min,k}^n + 6\epsilon(a_{\min} c_{\min})^{-1}$ and $n > (N_1 \lor N_2 \lor N_3 \lor N_\rho)$.

Now, we consider the second case where $g(\mu)$ is decreasing. As (A.14) is negative for all $\mu$, the minimizer of (A.13) is $\mu_{\min,k}^n = \mu_{\max,k}$ for all $n$. Following the same steps as in part 1, part 2 follows.

Now, we analyze the part 3. As $f(\cdot)$ is concave, $f'(n^{\alpha-1}C g(\mu)) \geq f'(n^{\alpha-1}C g(\mu_{\max}))$ for all $\mu_{\min} \leq \mu \leq \mu_{\max}$. Using this and convexity of $c(\mu)$, there exists an $N_4$, such that for $n > N_4$ the derivative in (A.14) is positive for all $\mu_{\min} \leq \mu \leq \mu_{\max}$, which in turn implies the maximizer of (A.13) is $\mu_{\min,k}^n = \hat{\mu}_{\max,k}^n$. Now, choosing $N_1$ such that $n > N_1$ implies $n^{\alpha-1} f'(n^{\alpha-1}C g(\mu_{\max})) > \frac{2a_{\min} c_{\min}}{c_{\min}}$ and following the same steps as in part 1, the theorem follows.

To prove Theorem 6.3 we need the following uniform integrability result:

**Lemma A.3.** For any $\alpha > 1/2$, the collection of random variables \( \{ (\hat{I}_n^\alpha(\infty) \cap (I^n(\infty) > 0))^{-1} \}_{n \in \mathbb{N}} \) is uniformly integrable.

**Proof.** We need to prove that for any $\rho > 0$, there exists an $M > 0$ such that
\[
 \sup_n \mathbb{E} \left[ \frac{n^\alpha}{I^n(\infty)} \mathbb{I}(I^n(\infty) > 0) \mathbb{I} \left( \frac{n^\alpha}{I^n(\infty)} > M \right) \right] < \rho. \tag{A.16}
\]
For any $n \in \mathbb{N}$, we have
\[
E \left[ \frac{n^\alpha}{I^n(\infty)} \mathbb{I}(I^n(\infty) > M) \right] = E \left[ \frac{n^\alpha}{I^n(\infty)} \mathbb{I}(I^n(\infty) > M) \mathbb{I} \left( \sum_{k=1}^{N^n} \mu_k^n - N^n \bar{\mu}_F \leq \frac{\beta}{2}(\lambda^n)^\alpha(\bar{\mu}_F)^{1-\alpha} \right) \right] + E \left[ \frac{n^\alpha}{I^n(\infty)} \mathbb{I}(I^n(\infty) > M) \mathbb{I} \left( \sum_{k=1}^{N^n} \mu_k^n - N^n \bar{\mu}_F > \frac{\beta}{2}(\lambda^n)^\alpha(\bar{\mu}_F)^{1-\alpha} \right) \right] \\
\leq E \left[ \frac{n^\alpha}{I^n(\infty)} \mathbb{I}(I^n(\infty) > M) \right] + n^\alpha \mathbb{P} \left( \left| \sum_{k=1}^{N^n} \mu_k^n - N^n \bar{\mu}_F \right| > \frac{\beta}{2}(\lambda^n)^\alpha(\bar{\mu}_F)^{1-\alpha} \right) \tag{A.17}
\]
First, we concentrate on the first term, on the right-hand side. From Assumption 3.1 for any $\epsilon_1 > 0$, there exists an $N_1$ such that for any $n > N_1$, $n^{-1}\lambda^n \leq \bar{\lambda} + \epsilon_1$. Hence, concentrating on $n > N_1$, we define
\[
K_\alpha = \left[ \frac{\beta(\bar{\lambda} + \epsilon_1)^\alpha}{4\mu_{\max}} \right]
\]
Now, we consider a sequence of birth-death processes $\{Y^n(t)\}_{n \in \mathbb{N}}$. The birth rate in the $n$th system is uniformly equal to $\lambda^n$ for any state. When the system is at state $Y^n(t) = i$, the death rate is given by
\[
\nu_i = \begin{cases} 
   i\mu_{\min} & \text{if } i \leq n \chi - n^\alpha K_\alpha \\
   \lambda^n + \frac{\beta}{4}(\lambda^n)^\alpha \bar{\mu}_F^{1-\alpha} & \text{if } i > n \chi - n^\alpha K_\alpha.
\end{cases}
\]
Now, we show that the process $N^n - Y^n(t)$ is stochastically smaller than $\hat{I}^n(t)$. Remembering that due to non-idling property, we have $\hat{I}^n(t) = (\hat{X}_n(t))^-$, we can write
\[
\hat{I}^n = \left( \hat{X}^n(0) \right)^- + n^{-\alpha} \sum_{i=1}^{A^n(t)} \mathbb{I} \left( \left( \hat{X}^n(\theta^n_{S_i}) \right)^+ = 0 \right) \mathbb{I} \left( U^n_t \leq \tilde{p}^n_i(\theta^n_{S_i}) \right) - n^{-\alpha} \sum_{i=1}^{A^n(t)} \mathbb{I} \left( \left( \hat{X}^n(\theta^n_{A_i}) \right)^- > 0 \right).
\]
Similarly, we can couple the birth-death process with the idleness process as
\[
Y^n(t) = Y^n(0) - n^{-\alpha} \sum_{i=1}^{A^n(t)} \mathbb{I} \left( U^n_t \leq \tilde{p}^n_i(\theta^n_{S_i}) \right) + n^{-\alpha} \sum_{i=1}^{A^n(t)} \mathbb{I} \left( \hat{Y}^n_i(\theta^n_{A_i})^- > 0 \right),
\]
where $\tilde{p}^n_i(\theta^n_{S_i}) = \frac{\mu_i}{\sum_{k=1}^{N^n} \bar{\mu}_k^n}$. Suppose $N^n - Y^n(0) = I^n(0)$ and define $\vartheta^n = \inf \{ t : N^n - Y^n(t) > I^n(t) \}$. Then, with probability 1, $N^n - Y^n(\vartheta^-) = I^n(\vartheta^-)$. As for any state where $I^n(t) = i$, $\frac{\mu_i}{\sum_{k=1}^{N^n} \bar{\mu}_k^n} \leq \frac{\sum_{k=1}^{N^n} \mu_k(1-I^n_k(t))}{\sum_{k=1}^{N^n} \mu_k}$, if $\vartheta$ is an event epoch for $S^n(t)$, we have $N^n - Y^n(\vartheta^-) \leq N^n - Y^n(\vartheta^-)$. If $\vartheta$ is

an event epoch for $A^n$, then $N^n - Y^n(\varnothing) \leq N^n - Y^n(\varnothing)$. This leads to a contradiction and we conclude that $N^n - Y^n(t)$ is stochastically less than $I^n(t)$.

Now, as $\nu_i = \lambda^n$ for all $i \geq N^n - n^\alpha K_{n,\alpha}$, for each $n$, the birth-death process $Y^n(t)$ is positive recurrent and for all $i \geq N^n - n^\alpha K_{n,\alpha}$, we have

\[
\mathbb{P}(Y^n(\infty) = i) = \left( \frac{\lambda^n}{\lambda^n + \frac{\beta}{2} (\lambda^n)^\alpha \bar{\mu}_F^{1-\alpha}} \right)^{i - N^n + n^\alpha K_{n,\alpha}} \mathbb{P}(Y^n(\infty) = N^n - n^\alpha K_{n,\alpha}).
\]

Hence,

\[
n^\alpha \mathbb{P} \left( Y^n(\infty) \geq N^n - n^\alpha K_{n,\alpha} \right) \leq \left( \frac{\lambda^n}{\lambda^n + \frac{\beta}{2} (\lambda^n)^\alpha \bar{\mu}_F^{1-\alpha}} \right)^{n^\alpha K_{n,\alpha}/2} \left( \lambda^n + \frac{\beta}{2} (\lambda^n)^\alpha \bar{\mu}_F^{1-\alpha} \right) \left( \lambda^n + \frac{\beta}{2} (\lambda^n)^\alpha \bar{\mu}_F^{1-\alpha} \right)^{n^\alpha K_{n,\alpha}/2}.
\]

When $\alpha > 1/2$, we see that the first-term on the right-hand side approaches 0 exponentially, whereas the second term increases linearly as $n \to \infty$, which implies that there exists an $N_2$ where the first term on the right-hand side of (A.17) is less than $\rho/2$ for $n > N_2$.

To address the second term on the right-hand side of (A.17), choose $p$ to be the smallest even number such that $p^\alpha > p/2 + \alpha$. Then, using Markov’s inequality

\[
n^\alpha \mathbb{P} \left( \sum_{k=1}^n \mu^n_k - N^n \bar{\mu}_F > \frac{\beta}{2} (\lambda^n)^\alpha (\bar{\mu}_F)^{1-\alpha} \right) \leq \frac{n^\alpha \sum_{j=1}^{n^\alpha} (\lambda^n)^\alpha (\bar{\mu}_F)^{1-\alpha}}{\left( \frac{\beta}{2} (\lambda^n)^\alpha (\bar{\mu}_F)^{1-\alpha} \right)^p}.
\]

It is now easy to see that the numerator scales with $p/2 + \alpha$ where as the denominator scales with $p\alpha$. Hence, the right-hand side approaches 0 as $n \to \infty$ and there exists an $N_3$ such that $n > N_3$ implies

\[
n^\alpha \mathbb{P} \left( \sum_{k=1}^n \mu^n_k - N^n \bar{\mu}_F > \frac{\beta}{2} (\lambda^n)^\alpha (\bar{\mu}_F)^{1-\alpha} \right) < \rho/2.
\]

Choosing $M = \max \{2/K_{n,\alpha}, N_1^n, N_2^n, N_3^n \}$, (A.16) holds and our result follows. \[\square\]

**Proof of Theorem 6.3** We define $U^n_{S,k,i} = 1$ if the $i$th event epoch of $A^n(t)$ corresponds to an arrival directly route to server $k$ and 0 otherwise. Similarly, we define $U^n_{A,k,i} = 1$ if the event epoch of $S^n(t)$ corresponds to an actual service completion at server $k$. Then, we have the following balance equation:

\[
I^n_k(t) = I^n_k(0) + \sum_{i=1}^{S^n(t)} U^n_{S,k,i} - \sum_{i=1}^{A^n(t)} U^n_{A,k,i}
\]

\[
= I^n_k(0) + \sum_{i=1}^{S^n(t)} U^n_{S,k,i} - \mu_k \int_0^t (1 - I^n_k(s-))ds + \mu_k \int_0^t (1 - I^n_k(s-))ds
\]

\[
- \left( \sum_{i=1}^{A^n(t)} U^n_{A,k,i} - \frac{\lambda^n}{n} \int_0^t \frac{h(\mu^n_k) I^n_k(s-)}{\langle h, \psi^n_{1,s-} \rangle} ds \right) - \frac{\lambda^n}{n} \int_0^t \frac{h(\mu^n_k) I^n_k(s-)}{\langle h, \psi^n_{1,s-} \rangle} ds.
\]
The second and fourth terms on the right-hand side are Poisson martingales with initial value 0. Assuming that the system is in stationarity, taking the expectation of both sides and using Fubini’s theorem, we have

\[
\frac{\lambda^n h(\mu)}{n} \mathbb{E} \left[ \frac{I^n_k(\infty)}{(h, \psi))_{\infty}} | \mu_k = \mu \right] + \mu \mathbb{E}[I^n_k(\infty)|\mu_k = \mu] = \mu
\]

Theorem 5.5 implies that \((h, \bar{\psi})_{\infty} \overset{D}{=} (h, \bar{\psi}_{\infty})\) where the limit is deterministic. Using Lemma A.3 and the notation \(L_F = \lambda(h, \bar{\psi}_{\infty})^{-1}\), we can find an \(N_\epsilon\) such that \(n > N_\epsilon\) implies

\[
|L_F h(\mu) \mathbb{E}[I^n_k(\infty)|\mu_k = \mu] + \mu \mathbb{E}[I^n_k(\infty)|\mu_k = \mu]| < \epsilon.
\]

Re-arranging the equation, the desired uniform convergence result holds.

**Proof of Theorem 6.4** Theorem 9 in [GDWW16] ensures that all the idle-time order based policies have the same stationary distribution. Hence, it is enough to prove the theorem only for one idle-time order based policy for each \(\alpha\). Hence, the case when \(\alpha = 1\) follows from Theorem 6.3 taking \(h(\mu) = 1\) for all \(\mu_{\text{min}} \leq \mu \leq \mu_{\text{max}}\). To prove the result for \(1/2 \leq \alpha < 1\), we concentrate on the longest-idle-server-first policy and follow a similar approach to the proofs of Lemma 4 and Theorem 6 in [BQ22].

As in the proof of Lemma 3.3, we define the \(i\)th event epoch of the arrival process \(A^n(t)\) as \(\theta^n_{A,i}\) and the inter-arrival time between arrival \(i - 1\) and \(i\) as \(u^n_i = \theta^n_{A,i} - \theta^n_{A,i-1}\). As the arrival process is a Poisson process, \(u^n_i\) are independent exponential random variables with rate \(\lambda^n\). Similarly, we define the \(i\)th epoch of the potential service completion process, \(S^n_k(t)\), of server \(k\) as \(\theta^n_{S,k,i}\) and the inter-event time between \((i-1)\)st and \(i\)th epoch as \(\nu^n_{k,i} = \theta^n_{S,k,i} - \theta^n_{S,k,i-1}\) for all \(i \in \mathbb{N}\), where \(\nu^n_{k,i}\) are independent exponential random variables with rate \(\bar{\mu}^n_k\). We also denote the \(i\)th epoch of the actual service completion process, \(D^n_k(t)\), of server \(k\) as \(\theta^n_{S,k,i}\). We define \(\bar{\phi}^n_{k,i}\), the idling time of server \(k\) after the \(i\)th server completion as \(\bar{\phi}^n_{k,i} := \inf\{t - \bar{\theta}^n_{S,k,i} : I^n_k(t) = 0, t > \bar{\theta}^n_{S,k,i}\}\). For longest-idle-server-first, we have

\[
\bar{\phi}^n_{k,i} = \sum_{j=2}^{I^n(\bar{\theta}^n_{S,k,i})} u^n_{A(\bar{\theta}^n_{S,k,i})+j} + (u^n_{A(\theta^n_{S,k,i})+1} - \bar{\theta}^n_{S,k,i}).
\]

Motivated by this, for the \(i\)th event epoch of the potential service completion process \(S^n_k(t)\), we associate the potential idling time of server \(k\), \(\phi^n_{k,i}\), defined as

\[
\phi^n_{k,i} = \sum_{j=2}^{I^n(\theta^n_{S,k,i})} u^n_{A(\theta^n_{S,k,i})+j} + (u^n_{A(\theta^n_{S,k,i})+1} - \theta^n_{S,k,i}).
\]

We also define \(\phi^n_{\infty,k} := \inf\{t \geq 0 : I^n_k(t) = 0\}\) as the first time server \(k\) is busy and

\[
\phi^n_{\infty,k} = \sum_{j=1}^{I^n(0)} u^n_j.
\]
Similar to equation (14) in [BQ22], we can write

\[
\int_0^t I^n_k(s) ds = (\phi^n_{-k} \land t) + \sum_{i=1}^{D^n_k(t)} (\tilde{\phi}^n_{k,i} \land (t - \tilde{\theta}^n_{S,k,i}^+))
\]

\[
= (\phi^n_{-k} \land t) + \sum_{i=1}^{\tilde{D}^n_k(t)} \tilde{\phi}^n_{k,i} - \sum_{i=1}^{\tilde{D}^n_k(t)} (\tilde{\phi}^n_{k,i} - t + \tilde{\theta}^n_{S,k,i}^+) \tag{A.18}
\]

For the proofs in this section, we assume that the system starts in stationarity. Hence, for all \( t \geq 0 \), we have

\[
\mathbb{E}[\mathbb{I}[t^n_k(\infty)] = \mu] = \mathbb{E}[t^n_k(t)] = \mu_k^n = \mu.
\]

We need the following lemma to prove our theorem.

**Lemma A.4.** For any \( k \), \( t^n_k(\infty) \xrightarrow{P} 0 \) and \( \mathbb{E}[t^n_k(\infty)] = \mu_k^n = \mu \) as \( n \to \infty \).

**Proof.** We have

\[
\mathbb{E}\left[\int_0^t I^n_k(s) ds | \tilde{\mu}^n_k = \mu \right] = \int_0^t \mathbb{E}[I^n_k(s) | \tilde{\mu}^n_k = \mu] ds
\]

\[
= \mathbb{E}[\mathbb{I}[\phi^n_{-k} \land t] | \tilde{\mu}^n_k = \mu] + \mathbb{E}\left[\sum_{i=1}^{\tilde{D}^n_k(t)} \mathbb{I}[\tilde{\phi}^n_{k,i} \land (t - \tilde{\theta}^n_{S,k,i})] | \tilde{\mu}^n_k = \mu \right]
\]

\[
\leq \mathbb{E}[\mathbb{I}[\phi^n_{-k}] | \tilde{\mu}^n_k = \mu] + \mathbb{E}\left[\sum_{i=1}^{\tilde{S}^n_k(t)} \mathbb{I}[\phi^n_{k,i}] | \tilde{\mu}^n_k = \mu \right]
\]

\[
\leq \frac{\mathbb{E}[t^n(\infty)]}{\lambda^n} + \mathbb{E}\left[\sum_{i=1}^{\tilde{S}^n_k(t)} \mathbb{I}[\mathbb{I}[\Theta^n_{S,k,i} \leq t] (\sum_{j=2}^{t^n(\Theta^n_{S,k,i})} u^n_{A(\Theta^n_{S,k,i})} + u^n_{A(\Theta^n_{S,k,i})} + 1 - \Theta^n_{S,k,i})] | \tilde{\mu}^n_k = \mu \right]
\]

\[
\leq \frac{\mathbb{E}[t^n(\infty)]}{\lambda^n} + \mathbb{E}\left[\sum_{i=1}^{\tilde{S}^n_k(t)} \mathbb{I}[t^n(\Theta^n_{S,k,i}) - 1 + \Theta^n_{S,k,i}] | \tilde{\mu}^n_k = \mu \right]
\]

\[
\leq \frac{\mathbb{E}[t^n(\infty)]}{\lambda^n} + \frac{\mu t^n(\infty) + 1 | \tilde{\mu}^n_k = \mu}{\lambda^n} \to 0
\]

As \( \mu \leq \mu_{\text{max}} \), the convergence is uniform. The convergence in probability follows using Markov’s inequality, which concludes the proof. \( \square \)
Our stationarity assumption combined with (A.18) implies that

$$n^{1-\alpha}E[I^n_k(\infty)|\tilde{\mu}_k^n = \mu] = \frac{n^{1-\alpha}}{t} \int_0^t E[I^n_k(s)|\tilde{\mu}_k^n = \mu] ds$$

$$= \frac{n^{1-\alpha}}{t} E[(\phi^n_{-k} \wedge t)|\tilde{\mu}_k^n = \mu] + \frac{n^{1-\alpha}}{t} E \left[ \sum_{i=1}^{D^n_k(t)} \tilde{\phi}^n_{k,i} |\tilde{\mu}_k^n = \mu \right]$$

$$- \frac{n^{1-\alpha}}{t} \sum_{i=1}^{D^n_k(t)} \left( \tilde{\phi}^n_{k,i} - \bar{\theta}^n_{S,k,i} \right)^+ |\tilde{\mu}_k^n = \mu \right].$$

Hence, for all $t \geq 0$, we can bound $n^{1-\alpha}E[I^n_k(\infty)|\tilde{\mu}_k^n = \mu]$ as

$$n^{1-\alpha}E[I^n_k(\infty)|\tilde{\mu}_k^n = \mu] \geq \frac{n^{1-\alpha}}{t} \sum_{i=1}^{D^n_k(t)} \phi^n_i |\tilde{\mu}_k^n = \mu - \frac{n^{1-\alpha}}{t} \sum_{i=1}^{D^n_k(t)} \left( \phi^n_i - (\tilde{\theta}^n_{S,k,i})^+ \right) |\tilde{\mu}_k^n = \mu \right]$$

$$n^{1-\alpha}E[I^n_k(\infty)|\tilde{\mu}_k^n = \mu] \leq \frac{n^{1-\alpha}}{t} \sum_{i=1}^{D^n_k(t)} \phi^n_i |\tilde{\mu}_k^n = \mu + \frac{n^{1-\alpha}}{t} E[\phi^n_{-k} |\tilde{\mu}_k^n = \mu]. \quad (A.19)$$

The first terms on the right-hand sides of (A.19) and (A.20) are the same and we first concentrate on this.

$$\frac{n^{1-\alpha}}{t} \sum_{i=1}^{D^n_k(t)} \phi^n_i |\tilde{\mu}_k^n = \mu = \frac{n^{1-\alpha}}{t} \sum_{i=1}^{D^n_k(t)} S^n_k(t)^n(\tilde{\theta}^n_{S,k,i}) + \int I^n_k(\theta^n_{S,k,i}) \sum_{j=1}^{I^n_k(\theta^n_{S,k,i})} u^n_{A(\theta^n_{S,k,i})+j} |\tilde{\mu}_k^n = \mu \right]$$

$$= \frac{n^{1-\alpha}}{t} \sum_{i=1}^{D^n_k(t)} S^n_k(t)^n(\tilde{\theta}^n_{S,k,i}) \sum_{j=1}^{I^n_k(\theta^n_{S,k,i})} u^n_{A(\theta^n_{S,k,i})+j} |\tilde{\mu}_k^n = \mu \right]$$

$$= \frac{n^{1-\alpha}}{t} \sum_{i=1}^{D^n_k(t)} S^n_k(t)^n(\tilde{\theta}^n_{S,k,i}) \sum_{j=1}^{I^n_k(\theta^n_{S,k,i})} u^n_{A(\theta^n_{S,k,i})+j} |\tilde{\mu}_k^n = \mu \right]$$

$$- \frac{n^{1-\alpha}}{t} \sum_{i=1}^{D^n_k(t)} I^n_k(\theta^n_{S,k,i}) \sum_{j=1}^{I^n_k(\theta^n_{S,k,i})} u^n_{A(\theta^n_{S,k,i})+j} |\tilde{\mu}_k^n = \mu \right]$$

$$= \frac{n^{1-\alpha}}{t} \left[ \int_0^t I^n_k(s) \sum_{j=1}^{I^n_k(s)} u^n_{A(\theta^n_{S,k,i})+j} dS^n_k(s) |\tilde{\mu}_k^n = \mu \right]$$

$$- \frac{n^{1-\alpha}}{t} \left[ \int_0^t I^n_k(s) \sum_{j=1}^{I^n_k(s)} u^n_{A(\theta^n_{S,k,i})+j} dS^n_k(s) |\tilde{\mu}_k^n = \mu \right]$$

$$= \frac{n^{1-\alpha}}{t} \left[ \int_0^t (\tilde{I}^n_k(s) + n^{-\alpha}) ds |\tilde{\mu}_k^n = \mu \right]$$

$$- \frac{n^{1-\alpha}}{t} \left[ \int_0^t I^n_k(s) (\tilde{I}^n_k(s) + n^{-\alpha}) ds |\tilde{\mu}_k^n = \mu \right].$$
Using the stochastic boundedness of $\hat{I}_n$ and Lemma A.4, the second term converges to 0. Using the stationarity assumption,

$$\sup_{\mu_{\text{min}} \leq \mu \leq \mu_{\text{max}}} \left| n \frac{\mu}{\lambda} t \mathbb{E} \left[ \int_0^t (\hat{I}_n^\alpha(s) + n^{-\alpha}) ds | \tilde{\mu}_k^n = \mu \right] \right| \to 0.$$  

Next, we concentrate on the second term on the right-hand side of (A.19). Each summand corresponds to the remaining idling time from the service completion at $\tilde{\theta}_{S,k,i}^n$. This implies only the summand corresponding to the last service completion is positive and

$$(\phi_i^n - (t - \tilde{\theta}_{S,k,i}^n))^+ \leq \sum_{j=1}^{I^n(t)} u_{A^n(t)+j}.$$  

Hence,

$$\frac{n^{1-\alpha}}{t} \mathbb{E} \left[ \sum_{i=1}^{D^n(t)} (\phi_i^n - (t - \tilde{\theta}_{S,k,i}^n))^+ | \mu_k^n = \mu \right] \leq \frac{n^{1-\alpha}}{t} \mathbb{E} \left[ \sum_{j=1}^{I^n(t)} u_{A^n(t)+j} | \mu_k^n = \mu \right]$$

$$= \frac{n^{1-\alpha}}{t} \mathbb{E} \left[ I^n(\infty) | \mu_k^n = \mu \right] = \frac{n}{\lambda^n t} \mathbb{E} \left[ \hat{I}_n^\alpha(\infty) | \mu_k^n = \mu \right].$$

Similarly, we can bound the second term on the right-hand side of (A.20) as

$$\frac{n^{1-\alpha}}{t} \mathbb{E} \left[ \phi_i^n | \mu_k^n = \mu \right] \leq \frac{n^{1-\alpha}}{t} \mathbb{E} \left[ \sum_{j=1}^{I^n(\infty)} u_j^n | \mu_k^n = \mu \right]$$

$$= \frac{n}{\lambda^n t} \mathbb{E} \left[ \hat{I}_n^\alpha(\infty) | \mu_k^n = \mu \right].$$

Taking $t$ large enough, both terms can be made arbitrarily small and hence (A.19) and (A.20) implies

$$\sup_{\mu_{\text{min}} \leq \mu \leq \mu_{\text{max}}} \left| n^{1-\alpha} \mathbb{E}[I_k^n(\infty) | \tilde{\mu}_k^n = \mu] - \frac{\mu}{\lambda} \mathbb{E}[I_\alpha(\infty) | \tilde{\mu}_k^n = \mu] \right| \to 0.$$  

Now, plugging in the expected value to be $\beta \lambda^{\alpha} \tilde{\mu}_F^{1-\alpha}(t, \eta_{\alpha,\infty})$ with $(t, \eta_{\alpha,\infty}) = \tilde{\bar{\mu}}_F^{-1}(\sigma_F^2 + \bar{\mu}_F^2)$, the result follows.  

**Proof of Lemma 6.6** The utility of idleness function is increasing and concave. Hence, to prove the concavity of the first part, it is enough to prove that the idleness experienced by a server with rate $\mu$, $(1 + L_{F_0} \hat{h}(\mu))^{-1}$, is concave (see e.g., page 86 in [BV04]). Taking the second derivative

$$\frac{d^2}{d\mu^2} (1 + L_{F_0} \hat{h}(\mu))^{-1} = \frac{L_{F_0}(L_{F_0}(2\hat{h}'(\mu)^2 - \hat{h}(\mu)\hat{h}''(\mu)) - \hat{h}''(\mu))}{(1 + L_{F_0} \hat{h}(\mu))^3}.$$  

Using convexity of $\hat{h}(\mu)$, we can conclude that the second derivative is negative for the idleness experienced by a server with rate $\mu$ is concave under (6.7) and hence, the result follows.
Proof of Theorem 6.8 As $L_F$ uniquely characterizes the distribution of idleness among servers, our result will follow if the distribution of service rates resulting from $L_F$ also yields the same $L_F$ for the $h$-random policy under consideration. Plugging in (5.6), this can be written as

$$\int_{\mu_{\min}}^{\mu_{\max}} \frac{1 + \beta}{\bar{\mu}F(1 + L_F h(\mu))} dF(\mu|L_F) = \beta.$$ 

Taking $\bar{\mu}F$ to the right-hand side and writing it as an integral, we get

$$\int_{\mu_{\min}}^{\mu_{\max}} \frac{1 + \beta}{1 + L_F h(\mu)} dF(\mu) = \beta \int_{\mu_{\min}}^{\mu_{\max}} \mu dF(\mu|L_F).$$

Re-arranging the terms, we get the desired result. □

Proof of Proposition 6.9 As $a_k^n$ is a continuous random variable, using the continuity of $C(\mu, L_{F_0})$ we can conclude that if $L_i \to L_{F_0}$, we have $F(\mu|L_i) \to F(\mu|L_{F_0})$ for all $\mu \in [\mu_{\min}, \mu_{\max}]$. Hence,

$$\left| \int_{\mu_{\min}}^{\mu_{\max}} \frac{1 - \beta L_i \tilde{h}(\mu)}{1 + L_i \tilde{h}(\mu)} dF(\mu|L_i) - \int_{\mu_{\min}}^{\mu_{\max}} \frac{1 - \beta L_{F_0} \tilde{h}(\mu)}{1 + L_{F_0} \tilde{h}(\mu)} dF(\mu|L_{F_0}) \right| \\
\leq \left| \int_{\mu_{\min}}^{\mu_{\max}} \frac{1 - \beta L_{F_0} \tilde{h}(\mu)}{1 + L_{F_0} \tilde{h}(\mu)} dF(\mu|L_i) - \int_{\mu_{\min}}^{\mu_{\max}} \frac{1 - \beta L_{F_0} \tilde{h}(\mu)}{1 + L_{F_0} \tilde{h}(\mu)} dF(\mu|L_{F_0}) \right| \\
+ \left| \int_{\mu_{\min}}^{\mu_{\max}} \frac{1 - \beta L_{F_0} \tilde{h}(\mu)}{1 + L_{F_0} \tilde{h}(\mu)} dF(\mu|L_i) - \int_{\mu_{\min}}^{\mu_{\max}} \frac{1 - \beta L_{F_0} \tilde{h}(\mu)}{1 + L_{F_0} \tilde{h}(\mu)} dF(\mu|L_{F_0}) \right|. $$

As $i \to \infty$, the first term on the right-hand side converges due to continuity of the integrand, and the second term converges using the definition of weak convergence, which implies the left-hand side of (6.11) is continuous with respect to $L_F$. Now, if $L_F = 1/(\beta \tilde{h}(\mu_{\min}))$, the integrand of (6.11) is non-negative for all $\mu \in [\mu_{\min}, \mu_{\max}]$ and $\tilde{h}(\mu)$ being strictly decreasing, implies that the integral is positive for all $L_F \leq 1/(\beta \tilde{h}(\mu_{\min}))$. Similarly, if $L_F = 1/(\beta \tilde{h}(\mu_{\max}))$, the integrand is non-positive and the integral is negative. Using the intermediate value theorem, the result follows. □

References

[AG10] G Allon and I Gurvich, Pricing and dimensioning competing large-scale service providers, Manufacturing and Service Operations Management 12 (2010), no. 3, 449–469.

[Arm05] M Armony, Dynamic routing in large-scale service systems with heterogeneous servers, Queueing Systems 51 (2005), no. 3-4, 287–329.

[ARS21] M. Armony, G. Roels, and H. Song, Capacity choice game in a multi-server queue: Existence of a nash equilibrium, Naval Research Logistics 68 (2021), no. 5, 663–678.

[ASS11] R. Atar, Y. Y. Shaki, and A. Shwartz, A blind policy for equalizing cumulative idleness, Queueing Systems 67 (2011), 275–293.

[Ata08] R Atar, Central Limit Theorem for a Many-Server Queue with Random Service Rates, The Annals of Applied Probability 18 (2008), no. 4, 1548–1568.

[AW10] M. Armony and A. R. Ward, Fair dynamic routing in large-scale heterogeneous-service systems, Operations Research 3 (2010), no. 58, 624–637.

[BBC19] E Bayraktar, A Budhiraja, and A Cohen, Rate Control Under Heavy Traffic with Strategic Servers, The Annals of Applied Probability 29 (2019), no. 1, 1–35.

[Bil99] P. Billingsley, Convergence of probability measures, Second, Wiley Series in Probability and Statistics: Probability and Statistics, John Wiley & Sons Inc., New York, 1999.

40
[BQ22] B. Büke and W. Qin, *Many-server queues with random service rates in the Halfin-Whitt regime: A measure-valued process approach*, Forthcoming in Mathematics of Operations Research, Available at Arxiv, https://arxiv.org/abs/1905.04198 (2022).

[BV04] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.

[Cal17] CallCentreHelper.com, *7 Tricks That Call Centre Employees Play*, 2017. https://www.callcentrehelper.com/7-tricks-that-call-centre-employees-play-67004.htm [Accessed 7-November-2022].

[CCS01] Y. Cohen-Charash and P. E. Spector, *The role of justice in organizations: A meta-analysis*, Organizational Behavior and Human Decision Processes 86 (2001), no. 2, 278–321.

[CCW01] J. A. Colquitt, D. E. Colon, M. J. Wesson, C. O. L. H. Porter, and K. Y. Ng, *Justice at the millennium: A meta-analytic review of 25 years of organizational justice research*, Journal of Applied Psychology 86 (2001), no. 3, 425–445.

[CH02] G. Cachon and P. Harker, *Competition and outsourcing with scale economies*, Management Science 48 (2002), no. 10, 1314–1333.

[CZ07] G. Cachon and F Zhang, *Obtaining fast service in a queueing system via performance-based allocation of demand*, Management Science 53 (2007), no. 3, 408–420.

[Del17] Deloitte, *Global Contact Center Survey*, Deloitte, 2017.

[GDWW16] R. Gopalakrishnan, S. Doroudi, A. R. Ward, and A. Wierman, *Routing and staffing when servers are strategic*, Operations Research 64 (2016), no. 4, 1033–1050.

[GKM03] N Gans, G Koole, and A Mandelbaum, *Telephone call centers: Tutorial, review, and research prospects*, Manufacturing and Service Operations Management 5 (2003), no. 2, 79–141.

[GLM+10] N Gans, N Liu, A Mandelbaum, H Shen, and H Ye, *Service times in call centers: Agent heterogeneity and learning with some operational consequences*, Borrowing strength: Theory powering applications, 2010, pp. 99–123.

[GMR02] O Garnett, A Mandelbaum, and M. Reiman, *Designing a Call Center with Impatient Customers*, Manufacturing & Service Operations Management 4 (2002), no. 3, 208–227.

[Go421] Go4Customer, *What is Call Avoidance?*, 2021. https://go4customer.co.uk/glossary/c/what-is-call-avoidance [Accessed 7-November-2022].

[Gop21] R. Gopalakrishnan, *On the architecture of service systems when servers are strategic*, 2021. Working paper, Smith School of Business, Queen’s University.

[Gum60] H Gumbel, *Waiting lines with heterogeneous servers*, Operations Research 8 (1960), no. 4, 504–511.

[GW09] I. Gurvich and W. Whitt, *Queue-and-idleness-ratio controls in many-server service systems*, Mathematics of Operations Research 34 (2009), 363–396.

[GW98] S. Gilbert and Z. Weng, *Incentive Effects Favor Nonconsolidating Queues in a Service System: The Principal-Agent Perspective*, Management Science 44 (1998), no. 12, 1662–1669.

[Has20] R. Hassin, *Rational queueing*, CRC Press, 2020.

[HH03] R. Hassin and M. Haviv, *To queue or not to queue: Equilibrium behavior in queueing systems*, Kluwer, 2003.

[HW81] S Halfin and H Whitt, *Heavy-Traffic Limits for Queues with Many Exponential Servers*, Operations Research 29 (1981), no. 3, 567–588.

[Jak86] A. Jakubowski, *On the Skorokhod topology*, Annales de l’institut Henri Poincare (B) Probability and Statistics 22 (1986), no. 3, 263–285.

[KKR92] E Kalai, M. Kamien, and M Rubinovitch, *Optimal Service Speeds in a Competitive Environment*, Management Science 38 (1992), no. 8, 1154–1163.

[PTW07] G. Pang, R. Talreja, and W. Whitt, *Martingale proofs of many-server heavy-traffic limits for Markovian queues*, Probability Surveys 4 (2007), 193–267.

[PWC18] PWC, *Experience is everything: Here’s how to get it right*, PWC, 2018. pwc.com/future-of-cx [Accessed 7-November-2022].

[RS15] J. Reed and Y. Shaki, *A fair policy for the G/G1/N Queue with Multiple Pools*, Mathematics of Operations Research 40 (2015), no. 3, 558–595.

[Sta20] Statista Research Department, *Global call center market size 2020-2027*, Statista, 2020.
[WA13] A. R. Ward and M. Armony, *Blind fair routing in large-scale service systems with heterogeneous customers and servers*, Operations Research 1 (2013), no. 61, 228–243.

[ZGW21] Y. Zhong, R. Gopalakrishnan, and A. Ward, *Behavior-aware queueing: The finite-buffer setting with many strategic servers*, 2021. submitted to Operations Research.

[ZW19] D. Zhan and A. R Ward, *Staffing, routing, and payment to trade off speed and quality in large service systems*, Operations Research (2019).