Optimal Routing to Parallel Servers in Heavy Traffic

Heng-Qing Ye *

Faculty of Business, Hong Kong Polytechnic University
Hung Hom, Hong Kong
lgtyehq@polyu.edu.hk

June 2023

Abstract

We study a system with heterogeneous parallel servers, each with an infinite waiting room. Upon arrival, a job is routed to the queue of one of the servers, possibly depending on the dynamic state information such as the real-time queue lengths, the arrival and service history of jobs. The objective is to find the routing policy that best utilizes the available state information to minimize the expected stationary queue length. In this paper, we establish the diffusion limit for the round-robin policy (resp. arrival-chasing policy, service-chasing policy), and show that with properly chosen parameters, it achieves the optimal performance asymptotically within the class of admissible policies that require no state information (resp. require arrival history, service history). Like the join-the-shortest-queue and the balanced routing policies that use real-time queue length information, the optimal service-chasing policy is also asymptotically optimal over all admissible policies. Further analysis of the diffusion limits yields a number of insights into the performance of these routing policies and reveals the value of various state information. We numerically demonstrate the effectiveness of the estimators derived from the diffusion limits for the policies being studied and obtain interesting observations. We also address the problem of interchange of limits under the aforementioned policies, which justifies the stationary performance of the diffusion limit as a valid approximation to that of the original system under respective policies. Methodologically, this study contributes to the application of the BIGSTEP method for constructing control policy to optimize stationary performance and the recipe for justifying the interchange of limits in the heavy-traffic analysis of stochastic processing networks.

Keywords: parallel server system, routing control, round robin, arrival chasing, service chasing, heavy traffic analysis.

1 Introduction

Routing control is an important component in many engineering and management systems, such as allocating processing orders in a machine shop, assigning cases to a panel of judges, or a web server routing requests to back-end servers. In this paper, we study a queueing system with parallel servers, each with an infinite waiting room, as depicted in Figure 1. One stream of jobs arrives at the system following a renewal process, and each job upon its arrival is routed immediately to one of the servers. At each server, the jobs are served according to the first-in first-out discipline, and their service times are independent and identically distributed (i.i.d.). The servers are heterogeneous and may have different service rates and service time distributions.

Imagine that upon the arrival of each job, a controller will evaluate the available (dynamic) state information and make a decision to dispatch the job to one of the servers. The state

*Supported in part by grant 15501421 and N_PolyU590/22 from RGC (HK).

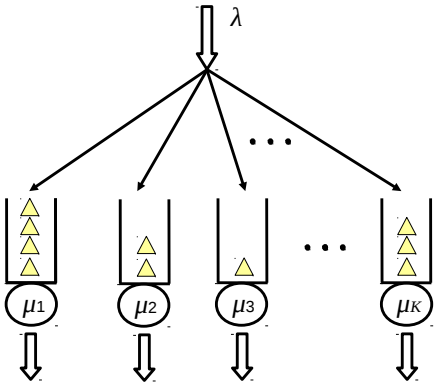

Figure 1: Network with parallel servers

information can be queue length, arrival history, service history, etc., depending on the nature of the application. For example, it would be costly for the domain name server (the router) to communicate intensively with the geographically distributed web servers to retrieve all the state information. Thus, the router may use only the information about the arrival history of requests (jobs) for routing control, or simply cyclically forwards the requests to servers. In this study, we consider non-anticipating routing policies only; that is, future information, such as the service time of a job before its service completion, will not be available for routing decision-making. We aim to identify the optimal routing policy that will dynamically employ the available state information to minimize the expected stationary (total) queue length, or equivalently, the average waiting time an arriving job experiences in steady state.

It is well-known that the join-the-shortest-queue (JSQ) and the balanced routing (BR) policies achieve the optimal performance asymptotically when the queue length state is observable (e.g. [16]). However, it would be useful to know what the optimal routing policy would be when the controller cannot observe any state of the system. Furthermore, we also want to understand the optimal policy when only a part of the system's state information can be observed, such as only the arrival history or service history. Answering these questions will help to reveal the value of various state information for routing control, which is particularly valuable for modern applications that are becoming more and more complicated and as new control mechanisms are being examined (e.g., [3, 54]). However, the dynamic nature of the system being studied, coupled with the complicated routing mechanism and arrival and service patterns, make it very difficult to answer these questions through exact analysis. To overcome this difficulty we apply the heavy-traffic analysis, which is an asymptotic approach that has been used to study a broad range of stochastic processing network systems.

We carry out the heavy-traffic analysis in two parts. In the first part, we establish the diffusion limit for a sequence of systems in heavy traffic, and use the limit to derive the optimal routing policy and the performance estimation heuristically. This is done in the spirit of the BIGSTEP method suggested by Harrison [30, 31]. For our problem, it includes formulating the diffusion limit for admissible routing policies, proposing the routing policy that optimizes the performance of the original system by examining the limit, and establishing diffusion limit theorem for the proposed policy to demonstrate its optimality and analyze its performance.

To illustrate the idea more formally, let $Q(t)$, a vector process, denote the queue length at time t in the original parallel server system. For technical and conceptual reasons, we consider an infinite sequence of copies or variations of the original system, indexed by n. Hence, let $Q^n(t)$ denote the queue length associated with the n-th system in the sequence; and let $\hat{Q}^n(t) := Q^n(n^2t)/n$ denote its diffusion-scaled version.

First, with the mindset that a practical routing policy must drive the system to certain

stationarity, we identify a class of admissible policies that are non-anticipating and will induce a diffusion limit for the sequence of systems in heavy traffic (class $\mathcal{D}$ and Proposition 1). We will see that the admissible class contains a wide range of policies of interest, such as the JSQ and BR policies, and yet are amenable to analysis. Focusing on this class, we study the routing control when certain state information is available for making routing control decisions.

Take the case that there is no dynamic state information available for routing upon each job arrival as an example (mainly, Section 4). We then identify the best possible limit $\hat{Q}(t)$ that involves no state information in routing. This is done through investigating the stationary performance of the general diffusion limit (for class $\mathcal{D}$) with the condition on that there is no relationship between the routing component and the arrival and service processes. It turns out that we interpret a blind routing policy, which is indeed a generalized round-robin (RR) policy, from the limit. Next, we justify the interpretation by establishing the diffusion limit theorem (Theorem 3). That is, under the RR policy, the sequence of systems does converge to the limit we identify, or $\hat{Q}^n(t) \Rightarrow \hat{Q}(t)$. Finally, the stationary performance of the limit, $\hat{Q}(\infty)$, is taken as an approximation of the stationary performance of the (original, discrete) system of interest, $\hat{Q}^n(\infty)$, which yields an approximation of the performance objective, i.e., the expected stationary queue length, immediately (cf. the estimator derived in (106)). Using the diffusion limit, we are able to gain useful insights into the RR policy too (cf. Subsection 4.1).

In a similar way, we establish and analyze an arrival-chasing (AC) policy and a service-chasing (SC) policy, which are asymptotically optimal when the controller can use the arrival history and service history information, respectively.

So far, we have followed a conventional approach to approximate the stationary performance $\hat{Q}^n(\infty)$ using $\hat{Q}(\infty)$, based on the pathwise diffusion limit ($\hat{Q}^n(t) \Rightarrow \hat{Q}(t)$) heuristically. Thus, in the second part of our heavy-traffic analysis, we provide a more rigorous justification for such a heuristics by establishing the convergence of stationary performance, $\hat{Q}^n(\infty) \Rightarrow \hat{Q}(\infty)$. This leads to the study of the famous interchange-of-limits problem, and we apply the recipe developed in Ye and Yao [66,67] to address the problem.

The crucial step in such a recipe is to bound the p-th moment of the state process ($p > m+1$, when the convergence of the m-th moment of the queue length is required). To do so, we establish a pathwise bound for the queue length process, known as the bounded workload condition in [66], under the RR, AC, SC and JSQ policies, respectively. It is a condition that the queue length can be bounded by a "free process" plus the initial queue length. Whereas for the BR policy, we do so by requiring a higher moment (p^*-th moment, with $p^* > 2(p+2)$) condition on the primitive arrival and service processes, as in [67]. Once the p-th moment is established, along with the uniform stability property, we can establish all other required properties leading to the convergence of stationary distribution and performance (cf. Theorems 8 and 9).

The contribution of this study is two-fold. First, in modeling analysis, we have formulated asymptotic optimal routing policies that utilize various kinds of state information. Using diffusion limits, we also analyze their performances and reveal the value of various state information as follows:

1. When there is no state information available for routing control, we establish that an RR policy is asymptotically optimal and derive its performance (Theorem 3). In the optimal RR policy, the routing rate (the portion of jobs dispatched to each server in the long run) exhibits a certain form of the square-root rule.

Examining the performance of the diffusion limit under the RR policy reveals interesting insights. The optimal RR policy can attain the performance of the JSQ policy (the globally optimal policy under heavy traffic) if and only if all service times are deterministic, and in this case, jobs are routed to each server proportional to the service rate. On the other hand, it could perform arbitrarily worse than the JSQ policy, say, when there are many servers in the system.

A conventional option in the class of RR policies, the proportional RR policy, is generally suboptimal within the class of RR policies, and can be arbitrarily worse than the optimal RR

policy. It coincides with the optimal RR policy if and only if variance-to-mean ratios of service times for all servers are the same.

2. When there is no state information and in addition the control is restricted to the class of probabilistic proportional (PP) routing policies, we identify and characterize the optimal one over this class (Theorem 4). In this optimal PP policy, the routing rate also exhibits certain form of square-root rule. The optimal PP policy performs strictly, and can even be arbitrarily, worse than the optimal RR policy.

3. When the arrival history information is available, we formulate a class of AC policies and show that an AC policy, with properly chosen parameters, is asymptotically optimal over all policies that utilize arrival history only (Theorem 5). In the optimal AC policy, the routing rate exhibits certain form of square-root rule once again.

Compared with the JSQ policy, the optimal AC policy, using only arrival history information, can achieve the JSQ performance when all servers, except at most one, have deterministic service times.

Comparing the optimal AC policy and the optimal RR policy, we find that when the variance-to-mean ratio of service times at all servers are equal, both policies yield the same expected stationary queue length, and thus using the arrival information does not help to minimize the queue length. Nevertheless, compared with the optimal RR policy, the optimal AC policy performs better in general, and can reduce up to 50% of the expected stationary queue length by utilizing the arrival history information.

4. When the service history information is available, we formulate a class of SC policies and show that an SC policy, with properly chosen parameters, is asymptotically optimal over all admissible policies (Theorem 6). Hence, the optimal SC policy achieves the JSQ performance. Interestingly, the diffusion limit under the optimal SC policy presents a *postponed* state-space collapse property; that is, the queue lengths, starting with any value, will evolve simultaneously and proportional to the service rates after some finite time.

5. We derive the heavy-traffic estimators of the system performance under the RR, AC and SC policies from the corresponding diffusion limits, and demonstrate through numerical studies that the estimators approximate the simulated performance (the proxy of the theoretical performance) of the original discrete system closely as the traffic intensity gets close to one (Section 8).

6. We address the problem of interchange of limits under the RR, AC, SC, JSQ and BR policies (Theorems 8 and 9), which justifies the stationary performance of the diffusion limit as a valid approximation to that of the original (diffusion-scaled) system under respective policies.

Second, in methodology, we have generated new ideas to execute the BIGSTEP method, and enriched the recipe in [66, 67] for justifying the interchange of limits in heavy-traffic analysis:

1. We have presented a comprehensive example of applying the framework of the BIGSETP method for constructing control policy to optimize stationary performance. Furthermore, we provide a rigorous justification of (stationary) diffusion approximation via the study of the interchange-of-limits problem. We believe these extensions to the method can be carried over to similar studies.

2. In [66,67] we have developed a recipe for justifying the (steady-state) diffusion approximation for a wide range of stochastic processing networks, such as the multiclass queueing network and the resource-sharing network. In the current study, we demonstrate that an additional routing mechanism feature can be incorporated and handled using the recipe straightforwardly, by enclosing a chasing process that captures the routing information in the Markovian state descriptor. In addition, we have refined the approach to establishing a uniform continuity property (Lemma 29), an important step in the recipe. This refinement of the recipe is transferable to the study of the interchange-of-limits problem for other models.

A brief review of the related literature is in order. On the optimal routing for the parallel server system, most of the earlier works assume that the controller can observe the queue

lengths of all (or some) of the servers. With the assumption of Poisson arrival and homogeneous exponential service, Winston [61] established the optimality of the JSQ policy in minimizing the long-run average delay for customers. This result was then extended in Weber [56] to allow more general arrival and service processes, with the service time having a non-decreasing failure rate. Readers are referred to Whitt [57] and Gupta *et al.* [26] for the literature review on the earlier work of the optimal routing control. Reiman [47] and Zhang [68] proved the heavy-traffic limit theorem for the system with the JSQ policy, and Chen and Ye [16] obtained a similar result for the BR policy. These results imply the asymptotic optimality of the JSQ and BR policies.

Closely related to optimal control, the stability of the JSQ and BR policies is also well understood in the literature (cf. [10, 16, 38]). We also provide a proof of the stability of the parallel system under various routing policies using the fluid model approach. Note that these stability results are established with all model parameters estimated accurately and the routing policy executed correctly; otherwise, refer to [46] for a recent study on systems involving "routing errors." These stability results ensure the existence of the stationary distribution and the stationary moments of the system, according to Dai and Meyn [17, 19]. Hurtado-Lange and Maguluri [34] established the convergence of stationary distributions for the JSQ systems in heavy traffic (interchange of limits) using a transformation method. Their result is a special case of the one here (Theorem 8) since in their model the interarrival times have all moments and the service times are bounded.

Another line of research related to the JSQ policy (and its variants) is concerned with large-scale systems where the number of servers and the associated queues increases to infinity along the sequence of systems. And, to analyze such systems, it would be natural to adopt the many-servers' assymptotic regime; refer to, e.g., [6, 22, 23, 29, 45, 54].

In most parts of the paper (except the numerical studies), we take the JSQ policy as the "benchmark." Nevertheless, some other policies, for example, the shortest-expected-delay (SED) policy and the MinDrift policy, are also globally asymptotic optimal for our model under heavy traffic and thus closely related to our study. The SED policy, which routes an arriving job to the queue that has the shortest expected delay, is a generalization of the JSQ policy. As pointed out by Selen *et al.* [49], if in addition to the real-time queue length information, we can estimate the expected service times, then the natural choice is to route the jobs according to SED, rather than JSQ. Stolyar [53] introduced the MinDrift routing rule in a multi-class output-queued system where the customer service time depends on both the customer class and server, and showed that the rule is asymptotically optimal under a complete resource pooling condition. The JSQ and SED policies for our model, a single-class system, are actually a special case of such a MinDrift policy. Indeed, it would be interesting to extend some of our studies (e.g., the optimal round-robin control) to such a multi-class system.

When the controller cannot observe any state information, the JSQ policy or its variations is not applicable. In the case of identical servers, the conventional RR routing policy is widely used, which in its simplest form assigns the incoming jobs to each server equally in a rotating fashion. It is shown that such an RR routing policy minimizes the long-run average total queue length in the system over all blind routing policies (Hajek [28] and Altman *et al.* [1]). Refer to, e.g., [35, 41, 42, 55, 63] for more related studies. Using heavy-traffic analysis, Tsoukatos and Makowski [55] established asymptotic optimality of the RR policy for the identical parallel server system. In their paper, the class of admissible policies that require no state information and admit a certain diffusion limit is introduced to demonstrate the optimality of the RR policy over all blind routing policies. In our study, we focus on a similar class of admissible policies, which may depend on certain state information, in the framework of the BIGSTEP method. We not only extend the asymptotic optimality of the RR policy to the case of heterogeneous server system, but can also identify the asymptotically optimal policies when the information about the arrival or service history is available for routing control.

In the situation that the controller can enclose the arrival or service history for routing

control, to the best of our knowledge the optimal control problem is not addressed in the literature. As mentioned above, we have formulated the AC and SC policies and characterized their optimality in such a situation, which also reveals the value of arrival and service history information for routing control. These results fill a gap in the literature, and have the potential for real-life applications.

There has been a vast volume of literature on heavy traffic analysis for queueing systems, which is the major tool for the theoretical analysis in this paper. Readers may refer to Chen and Yao [15] for earlier literature on heavy-traffic analysis. Here we briefly highlight a few methods that play key roles in our paper.

In establishing the diffusion limit theorems (Theorems 5 and 6) and the convergence of stationary performance (Theorem 9), we have adopted the hydrodynamic approach of Bramson [9] and its variations (e.g., [43,52,53,64–67]). By this technique, the order $O(n^2)$-long time interval of the diffusion-scaled process is broken down into $O(n)$ pieces of $O(n)$-long time intervals, and thereafter the diffusion-scaled process is converted to $O(n)$ pieces of fluid-scaled "magnifiers" of the corresponding processes. Then the properties developed for the fluid-scaled process can be applied to establish the key properties of the diffusion-scaled process. This approach has been applied for establishing diffusion limit for many complex stochastic processing network systems; refer to the references just mentioned.

The BIGSTEP method we follow to identify the optimal policies was first suggested by Harrison [30, 31]. A distinct example for this method is presented in Bell and Williams [5] for the optimal scheduling of a two-server system in heavy traffic, where the optimality is justified via a diffusion limit theorem. More examples of using this method to study the optimal control of queueing networks can be found in, e.g., [2, 18, 33, 37, 44]. These studies are based on the pathwise minimality of the diffusion limit under the proposed control policy, whereas in this paper, we demonstrate a comprehensive example of studying the optimal control for stationary performance following the method.

Justifying the diffusion limit as the valid approximation of the stationary performance of the original system leads to the problem of interchange of limits. Refer to [66,67] and the references there for a detailed account on this subject. Recent studies on this problem have been initiated by Gamarnik and Zeevi [24] and Budhiraja and Lee [11] for the generalized Jackson network. These works take advantage of the Lipschitz continuity property of the Skorohod (reflection) mapping associated with the network system concerned, so as to convert the uniform moment bound of arrival and service processes to that of the key performance measures such as the queue length processes. The latter turns out to be the key property for addressing the problem. However, the Lipschitz continuity property is often difficult to establish or is even not available for many other models, and subsequent works (e.g., [25,66,67]) attempt to relax this requirement. Particular, the recipe we developed in [66,67] will be used to address the problem of interchange of limits for the parallel server system being studied. As mentioned above, the key is to verify the bounded queue length (or workload) condition or the p^*-th moment condition. These two conditions do not imply each other. To verify the bounded queue length condition requires effort, but once it is verified, the interchange is justified without requiring any higher order moments on the primitives. In contrast, the p^*-th moment condition is trivial to verify, and indeed automatically holds in networks where the primitives have moments of all orders (e.g., renewal arrivals with phase-type interarrival times and i.i.d. phase-type service times). The approach is applied to address the interchange-of-limits problem for a queueing system with customer abandonment recently (cf. [39, 40]) and for the parallel server system being studied.

The paper is organized as follows. The parallel server system is described in the next section. In Section 3, we introduce the class of admissible policies within which we will study the optimal routing policies. We establish the lower-bound performance of the admissible policy and show that the JSQ and BR policies are admissible and achieve the lower bound. In the next four sections, we formulate and characterize the asymptotically optimal routing policies. Specifically,

in Section 4, we identify the optimal RR policy when there is no state information available for routing control decision-making. We demonstrate its optimality by establishing a (process-wise) diffusion limit theorem, and use the diffusion limit to understand the performance of the RR policy. The PP policy, another class of policies that require no state information, are studied in Section 5. Following the similar approach, in Section 6 and 7, we study the optimal AC and SC policies that utilize the arrival history and service history respectively. Numerical studies for evaluating the heavy-traffic estimators are presented in Section 8, where preliminary observations about the performance of various policies for a non-heavily loaded system is also illustrated via a simulation study. While in the above the analysis of the stationary performance and optimality of various policies are built on process-wise diffusion limit theorems, Section 9 is devoted to the study of the interchange-of-limits problems. It justifies the stationary performance of the diffusion limit as a valid approximation of the stationary performance of the pre-limit systems. Some preliminaries about the Skorohod mapping and all proofs for main results are presented in the Appendix. (Section 9 and the Appendix are available in the Electronic Companion.)

2 Model and Preliminary

We consider a queueing system with $K(\geq 2)$ servers, indexed by $k \in \mathcal{K} = \{1, \cdots, K\}$, following [16]. Each server has an infinite waiting room. Jobs arrive at the system following a renewal process with arrival rate λ. Upon arrival, each job is routed to one of the queues to attain service. At server k, jobs are served at a rate of μ_k following the order of arrival.

Denote the interarrival time between consecutive arrivals by u_ℓ, $\ell = 1, 2, \cdots$. Denote the service time of ℓ-th job at server k by $v_{k,\ell}$, $\ell = 1, 2, \cdots$, and $k \in \mathcal{K}$. That is, the service time of a job is server-dependent. We assume that the interarrival time sequence $\{u_\ell, \ell \geq 1\}$, and the service time sequences $\{v_{k,\ell}, \ell \geq 1\}$, $k \in \mathcal{K}$, are mutually independent i.i.d. random sequence, all with finite second moments. In particular, let u_ℓ have mean $1/\lambda$ and coefficient of variation c_a, and let $v_{k,\ell}$ have mean $1/\mu_k$ and coefficient of variation $c_{b,k}$, $k \in \mathcal{K}$.

For ease of presentation, we denote the sum of a set of numbers and functions, say $\{x_i, i \in \mathcal{I}\}$ or $\{f_i(t), i \in \mathcal{I}\}$, as $x_{\mathcal{A}} = \sum_{i\in\mathcal{A}} x_i$ and $f_{\mathcal{A}}(t) = \sum_{i\in\mathcal{A}} f_i(t)$, respectively, for any $\mathcal{A} \subset \mathcal{I}$. The L_1-norm of a vector is then $|x| = \sum_{i\in\mathcal{I}} |x_i|$. These representations of summation and L_1-norm will be used interchangeably in this paper. Hence, $\mu_{\mathcal{K}} = \sum_{k\in\mathcal{K}} \mu_k$ denotes the total service rate of the system, and then $\rho = \lambda/\mu_{\mathcal{K}}$ is the (nominal) traffic intensity of the system.

Next, we introduce the following related processes:

$$\Upsilon(\ell) = \sum_{\ell'=1}^{\ell} u_{\ell'}, \quad E(t) = \sup\{\ell : \sum_{\ell'=1}^{\ell} u_{\ell'} \leq t\}, \quad S_k(t) = \sup\{\ell : \sum_{\ell'=1}^{\ell} v_{k,\ell'} \leq t\}.$$

We call $E(t)$, $t \geq 0$, the (exogenous) arrival process, which denotes the number of arrivals during the time interval $[0, t]$. $\Upsilon(\ell)$ records the arrival time of the ℓ-th job (for $\ell = 0, 1, 2, \cdots$, with $\Upsilon(0) = 0$), and hence can be referred to as the arrival time process. We call $S(t) = (S_k(t))_{k\in\mathcal{K}}$, $t \geq 0$, the service process, where $S_k(t)$ denotes the number of class-k service completions (job departures) after server k is busy for a total of t time units.

To describe the routing of jobs, we define routing sequence $\phi(\ell) = (\phi_k(\ell))_{k\in\mathcal{K}}$, $\ell = 0, 1, 2, \cdots$, as

$$\phi_k(\ell) = \begin{cases} 1 & \text{if the } \ell\text{-th arrival is routed to class-}k, \\ 0 & \text{otherwise.} \end{cases}$$

Let the routing process be $\Phi(\ell) = (\Phi_k(\ell))_{k\in\mathcal{K}}$, $\ell = 0, 1, 2, \cdots$, where $\Phi_k(\ell)$ is the number of jobs among the the first ℓ arrivals that are dispatched to the server k. The total number of jobs routed to servers must be equal to the total arrivals:

$$\Phi_k(\ell) = \sum_{\ell'=1}^{\ell} \phi_k(\ell'), \qquad \sum_{k\in\mathcal{K}} \Phi_k(\ell) = \ell. \tag{1}$$

We assume the routing policy is *non-anticipating*; that is, at each time t of a job arrival, the policy depends only on past history in a measurable way (cf. Williams [59], Section 3.1.5). For example, the controller cannot observe and hence is not allowed to utilize the service time of any job before its service completion.

We focus on the routing controls that will minimize the performance objective, expected stationary (total) queue length, i.e., $\mathsf{E}Q_{\mathcal{K}}(\infty)$ $(= \mathsf{E}\sum_{k\in\mathcal{K}} Q_k(\infty))$, where $Q_k(\infty)$ represents the stationary queue length of server k, if exists. Then, according to the Little's Law, the expected waiting time of jobs (the time between a job's arrival and its service completion in stead-state) can be derived by dividing $\mathsf{E}Q_{\mathcal{K}}(\infty)$ by the arrival rate λ. Therefore, minimizing the waiting time and total queue length in stead-state are equivalent.

Some routing policies that dictate the routing process $\Phi(\ell)$ are studied in this paper. Below, we describe JSQ, BR, RR and PP routing policies, which are both theoretically and practically important and are widely studied in the literature. The first two are state-dependent policies that use the real-time queue length information in dispatching new arrivals, while the latter two are blind policies that do not use any state information. Later, we introduce two additional routing policies for improving system performance: an AC policy and an SC policy, both of which are non-anticipating and utilize some real-time information about the arrival and the service processes respectively.

The JSQ policy dispatches a new arrival to the queue that has the shortest queue length. The BR policy with a given integer c $(2 \le c \le K)$ and a probability distribution $\pi = (\pi_1, \cdots, \pi_K)$ is described as follows. When a job arrives, c servers are first chosen sequentially — at each step, server k is chosen with probability π_k $(k \in \mathcal{K})$, independent of the choice made for the previously arrived jobs; repeat until c distinct servers are chosen. Then the job is routed to the server that has the shortest queue among the c chosen servers. Typically, choosing π in proportional to the service rate (i.e., $\pi_k = \mu_k/\mu_{\mathcal{K}}$) will yield an asymptotically optimal performance; and it is interesting to note that BR policy is robust subject to small change in the value of π (cf. [16]). By letting $c = K$, the BR policy is reduced to the JSQ policy.

In the above JSQ and BR policies, if there is a tie by having more than one shortest queue, the job can be routed to any one of the tied servers. Actually, the main results remain the same under any tie-breaking methods in our asymptotic analysis. Here and below, we omit the detail of tie-breaking methods in all the routing policies being studied.

Under the RR policy with the weight parameter (also a probability distribution) $p = (p_1, \cdots, p_K)$ satisfying $\sum_{k\in\mathcal{K}} p_k = 1$, a fraction p_k of jobs is sent to the server k according to a pre-specified splitting sequence. For example, when all the servers are treated equally, i.e., all p_k's are the same, the RR policy routes incoming jobs to servers in order and in rotating fashion; specifically, the $(\ell K+k)$th job is sent to the kth server, $k = 1, \cdots, K$ and $\ell = 0, 1, 2, \cdots$. Indeed, this conventional RR routing control is often applied to the identical server setting.

More generally, the sequence of arrivals should be split so that the number of jobs dispatched to each server k is "close" to its quota, a fraction p_k of the total arrival, at any time instance. More specifically, in addition to equation (1), the RR policy should satisfy the following requirement: for some constant κ,

$$|\Phi_k(\ell) - p_k\ell| < \kappa, \quad \ell = 1, 2, \cdots, \quad k \in \mathcal{K}. \tag{2}$$

Here, we have specified a condition that characterizes a (generalized) RR routing, which is sufficient for our purpose in the asymptotic analysis below. As a concrete example, an implementation is as follows. Upon the ℓ-th arrival $(\ell = 1, 2, \cdots)$, pick (one of) the server, denoted k', that has the smallest "surplus":

$$k' = \arg\min_{k\in\mathcal{K}} (\Phi_k(\ell - 1) - p_k\ell). \tag{3}$$

Then, send the ℓ-th job to server k', i.e., let

$$\Phi_{k'}(\ell) = \Phi_{k'}(\ell - 1) + 1, \quad \text{and} \quad \Phi_k(\ell) = \Phi_k(\ell - 1), k \neq k'.$$

The PP policy is also specified with the weight parameter $p = (p_1, \cdots, p_K)$. That is, upon the ℓ-th arrival, this job is dispatched to the server k with probability p_k,

$$\mathsf{P}\{\phi(\ell) = e^k\} = p_k, \tag{4}$$

where e^k a K-dimensional vector with its k-th component being one and other components being zero. The routing sequence $\{\phi(\ell), \ell = 1, 2, \cdots\}$ are mutually independent random vectors, and are also independent of all the interarrival times and service times. Note that, by allowing $c = 1$, the BR policy will become a PP policy.

Clearly, given the weight parameter p, the RR and PP policies use no dynamic state information and hence are "blind" policies. We will see later that to specify the weight (control) parameter optimally requires a priori knowledge of all model parameters, i.e., the first two moments of the interarrival and service times.

Here we describe the key performance measure of the system. Let $Q(t) = (Q_k(t))_{k\in\mathcal{K}}$ be the queue length at time t, where $Q_k(t)$ denotes the number of jobs in queue k at time t. Then, the number of arrivals routed to server k during $[0, t]$, is given as $\Phi_k(E(t))$, and satisfies the requirement in (1):

$$\sum_{k\in\mathcal{K}} \Phi_k(E(t)) = E(t). \tag{5}$$

Let $B(t) = (B_k(t))_{k\in\mathcal{K}}$, where $B_k(t)$ denotes the busy time, i.e., total amount of the time that server k has served jobs during $[0, t]$. With the busy time process, the number of service completions at server k up to time t is given as $S_k(B_k(t))$. Then, the dynamics of the queueing system is characterized by

$$Q_k(t) = Q_k(0) + \Phi_k(E(t)) - S_k(B_k(t)) \geq 0, \tag{6}$$

$$B_k(t) = \int_0^t 1_{\{Q_k(s)>0\}} ds. \tag{7}$$

The first equation is a balanced equation, assuming the initial queue length (at time 0) is $Q_k(0)$ for $k \in \mathcal{K}$. The second equation specifies a work-conserving condition, i.e., the server must work at its full capacity when there is at least one job in its queue. Define the idling processes $Y(t) = (Y_k(t))_{k\in\mathcal{K}}$ as follows,

$$Y_k(t) = \mu_k(t - B_k(t)) = \mu_k \int_0^t 1_{\{Q_k(s)=0\}} ds. \tag{8}$$

It is immediately observed from the above expressions that for $k \in \mathcal{K}$,

$$\int_0^\infty Q_k(s) dY_k(s) = 0, \tag{9}$$

$$Y_k(t) \text{ is non-decreasing in } t \geq 0, \text{ and } Y_k(0) = 0. \tag{10}$$

To carry out the heavy traffic analysis, we introduce a sequence of systems, indexed by $n \in \mathcal{N}$, where $\mathcal{N}$ can be chosen as the set of natural numbers, or more generally, a sequence of positive real numbers that increase to $+\infty$. Each system is like the one introduced above, but may differ in their arrival rates. Specifically, for the n-th system, the interarrival times and service times are denoted as u_ℓ^n and $v_{k,\ell}^n$, with the first and second order parameters (λ^n, c_a^n) and $(\mu_k^n, c_{b,k}^n)$, respectively. For convenience, we assume that the mean and variance of service times do not change with n, hence the index n is omitted for these parameters (i.e., $(\mu_k^n, c_{b,k}^n) \equiv (\mu_k, c_{b,k})$). The processes defined for the n-th system are then the arrival process $E^n(t)$, the arrival time sequence $\Upsilon^n(\ell)$, the service process $S^n(t) = (S_k^n(t))_{k\in\mathcal{K}}$, the routing sequence $\phi^n(\ell) = (\phi_k^n(\ell))_{k\in\mathcal{K}}$, the

routing process $\Phi^n(\ell) = (\Phi^n_k(\ell))_{k\in\mathcal{K}}$, the queue length process $Q^n(t) = (Q^n_k(t))_{k\in\mathcal{K}}$, the busy time process $B^n(t) = (B^n_k(t))_{k\in\mathcal{K}}$, and the idling process $Y^n(t) = (Y^n_k(t))_{k\in\mathcal{K}}$. These processes satisfy the relationships in equations (5-10), with the index n properly appended.

We assume the sequence of systems are linked through the limit,

$$\lambda^n \to \lambda := \mu_{\mathcal{K}} \text{ and } c_a^n \to c_a, \text{ as } n \to \infty,$$

and furthermore the following *heavy traffic condition* is satisfied,

$$n(\lambda^n - \mu_{\mathcal{K}}) \to \theta_{\mathcal{K}} < 0, \text{ as } n \to \infty. \tag{11}$$

From now on, the parameter λ denotes the limit of λ^n rather than the arrival rate of a particular system. Though λ takes on the same constant value as $\mu_{\mathcal{K}}$, it explicitly indicates that the value is arrived at through a convergence procedure; this is useful when we infer the heuristic policy for the pre-limit system from the limit theorems to be established below. Moreover, the above condition implies that the (nominal) traffic intensity approaches one,

$$\rho^n := \frac{\lambda^n}{\mu_{\mathcal{K}}} \to 1, \text{ as } n \to \infty.$$

Without loss of generality, we assume $\lambda^n < \mu_{\mathcal{K}}$, or $\rho^n < 1$, for all $n \in \mathcal{N}$. We also assume at least one of the coefficients of variation, c_a and $c_{b,k}$ ($k \in \mathcal{K}$), is positive; otherwise, the system becomes deterministic and is trivial in heavy traffic analysis.

In addition, to guarantee the convergence of the fluid-scaled and diffusion-scaled primitive processes below, we make the following standard assumption: there exists a function $g(a)$, satisfying $g(a) \to 0$ as $a \to \infty$, such that the following holds uniformly on n,

$$\mathsf{E}[(u^n_\ell)^2 \mathbf{1}\{u^n_\ell > a\}] \le g(a) \text{ and } \mathsf{E}[(v^n_{k,\ell})^2 \mathbf{1}\{v^n_{k,\ell} > a\}] \le g(a), \quad \text{for } k \in \mathcal{K}, \text{ and all } a \ge 0. \tag{12}$$

This assumption, first introduced by Bramson [9] (see also [16, 43, 52]), imposes control on the fluctuation of the arrival and service processes and thus the sample paths of the system state.

We apply the standard fluid scaling to the primitive processes associated with the sequence of systems:

$$\left(\bar{E}^n(t), \bar{S}^n(t), \bar{\Upsilon}^n(t)\right) := \frac{1}{n}\left(E^n(nt), S^n(nt), \Upsilon^n(\lfloor nt \rfloor)\right), \tag{13}$$

where the floor function $\lfloor x \rfloor$ gives the greatest integer less than or equal to x for any real number x. Similarly, define the fluid-scaled version of the derived processes:

$$\left(\bar{Q}^n(t), \bar{\Phi}^n(t), \bar{B}^n(t), \bar{Y}^n(t)\right) = \frac{1}{n}\left(Q^n(nt), \Phi^n(\lfloor nt \rfloor), B^n(nt), Y^n(nt)\right). \tag{14}$$

For the scaled primitive processes, when $n \to \infty$ and under the assumption (12), we have the following functional strong law of large numbers with probability one,

$$\left(\bar{E}^n(t), \bar{S}^n(t), \bar{\Upsilon}^n(t)\right) \to (\lambda t, \mu t, \lambda^{-1} t), \quad \text{u.o.c. of } t \ge 0. \tag{15}$$

This convergence is a direct consequence of Lemma 13 (in Appendix A.1), which is useful in heavy-traffic analysis.

In addition to the fluid scaling defined in (13,14), it would be convenient to introduce an alternative version by replacing the scaling factor n by n^2,

$$\begin{aligned} &\left(\tilde{E}^n(t), \tilde{S}^n(t), \tilde{\Upsilon}^n(t)\tilde{Q}^n(t), \tilde{\Phi}^n(t), \tilde{B}^n(t), \tilde{Y}^n(t)\right) \\ := \;& \frac{1}{n^2}\left(E^n(n^2 t), S^n(n^2 t), \Upsilon^n(\lfloor n^2 t \rfloor), Q^n(n^2 t), \Phi^n(\lfloor n^2 t \rfloor), B^n(n^2 t), Y^n(n^2 t)\right). \end{aligned}$$

Clearly, the following functional strong law of large numbers, just like (15), continues to hold with probability one,

$$\left(\tilde{E}^n(t), \tilde{S}^n(t), \tilde{\Upsilon}^n(t)\right) \to (\lambda t, \mu t, \lambda^{-1} t), \quad \text{u.o.c. of } t \geq 0. \tag{16}$$

Define the diffusion scaling (along with centering) for the primitive processes:

$$\left(\hat{E}^n(t), \hat{\Upsilon}^n(t), \hat{S}^n_k(t)\right) := \frac{1}{n}\left(E^n(n^2 t) - \lambda^n n^2 t, \Upsilon^n(\lfloor n^2 t \rfloor) - (\lambda^n)^{-1}\lfloor n^2 t \rfloor, S^n_k(n^2 t) - \mu_k n^2 t\right).$$

By the functional central limit theorem for the renewal process (see, for example, Chapter 5 of [15]), we have the following weak convergence,

$$(\hat{E}^n(t), \hat{\Upsilon}^n(t), \hat{S}^n(t)) \;\Rightarrow\; (\hat{E}(t), -\lambda^{-1}\hat{E}(\lambda^{-1}t), \hat{S}(t)), \qquad \text{as } n \to \infty, \tag{17}$$

where $\hat{E}(t)$ is a (one-dimensional) Brownian motion with zero mean and variance λc_a^2; and $\hat{S}(t) = (\hat{S}(t))_{k\in\mathcal{K}}$ is a K-dimensional Brownian motion with independent coordinates, whose kth coordinate, $\hat{S}_k(t)$, is a Brownian motion with zero mean and variance $\mu_k c_{b,k}^2$. $\hat{E}(t)$ and $\hat{S}(t)$ are independent.

For the routing process, we introduce a set of parameters $(p^n_k)_{k\in\mathcal{K}}$ satisfying $\sum_{k\in\mathcal{K}} p^n_k = 1$ and $p^n_k \geq 0$, and denote formally:

$$\hat{\Phi}^n_k(t) := \frac{1}{n}\left(\Phi^n_k(\lfloor n^2 t\rfloor) - p^n_k \lfloor n^2 t\rfloor\right) = \frac{1}{n}\sum_{\ell=1}^{\lfloor n^2 t\rfloor}\left(\phi^n_k(\ell) - p^n_k\right), \quad k \in \mathcal{K}. \tag{18}$$

A routing policy for the sequence of systems just introduced actually refers to a sequence of policies, with the n-th policy associated with the n-th system. As we will see below, when the routing policy and the parameters p^n_k are properly specified, the scaling in equation (18) is indeed a proper diffusion-scaling and satisfies a central limit theorem.

For the other derived processes, we write:

$$\hat{Q}^n(t) := \frac{1}{n}Q^n(n^2 t), \quad \hat{Y}^n(t) := \frac{1}{n}Y^n(n^2 t), \quad k \in \mathcal{K}. \tag{19}$$

Rewrite equation (6) for the n-th system as:

$$\begin{aligned} Q^n_k(t) &= Q^n_k(0) + \Phi^n_k(E^n(t)) - S^n_k(B^n_k(t)) \\ &= Q^n_k(0) + X^n_k(t) + Y^n_k(t), \qquad (20) \\ X^n_k(t) &= [\Phi^n_k(E^n(t)) - p^n_k E^n(t)] + p^n_k[E^n(t) - \lambda^n t] \\ &\quad -[S^n_k(B^n_k(t)) - \mu_k B^n_k(t)] + (p^n_k\lambda^n - \mu_k)t. \qquad (21) \end{aligned}$$

Then, the dynamics given in equations (5-10) can be written in diffusion scaling for the n-the system: for all $k \in \mathcal{K}$ and $t \geq 0$,

$$\hat{Q}^n_k(t) = \hat{Q}^n_k(0) + \hat{X}^n_k(t) + \hat{Y}^n_k(t) \geq 0, \tag{22}$$

$$\int_0^\infty \hat{Q}^n_k(s) d\hat{Y}^n_k(s) = 0, \tag{23}$$

$$\hat{Y}^n_k(t) \text{ is non-decreasing in } t \geq 0, \text{ and } \hat{Y}^n_k(0) = 0, \tag{24}$$

where

$$\hat{X}^n_k(t) = \hat{\Phi}^n_k(\tilde{E}^n(t)) + p^n_k\hat{E}^n(t) - \hat{S}^n_k(\tilde{B}^n_k(t)) + n(p^n_k\lambda^n - \mu_k)t, \tag{25}$$

$$\sum_{k\in\mathcal{K}} \hat{\Phi}^n_k(t) = 0. \tag{26}$$

Once the routing control $\hat{\Phi}^n(t)$ and the initial state $\hat{Q}^n(0)$ are specified, the above Skorohod problem will determine the queue length process $\hat{Q}^n(t)$ (cf. Lemma 10).

3 Admissible Control and Lower Bound

A routing policy for the sequence of systems introduced above, denoted as D, is called *diffusion-admissible*, or simply *admissible*, if they are non-anticipating and there exist positive constants $p^n = (p_k^n)_{k\in\mathcal{K}}$ for each $n \in \mathcal{N}$, satisfying $\sum_{k\in\mathcal{K}} p_k^n = 1$, such that the followings hold:

(a) For some constants $\theta = (\theta_k)_{k\in\mathcal{K}} < 0$, the following holds for $\theta^n = (\theta_k^n)_{k\in\mathcal{K}}$,

$$\theta_k^n := n(p_k^n\lambda^n - \mu_k) \to \theta_k, \quad \text{as } n \to \infty. \tag{27}$$

(b) We can find a set of states $\mathcal{G} \subset \mathbb{R}_+^K$ (the non-negative orthant of the K-dimensional real space), which serves as "good" initial (limiting) states and typically includes the origin. (We write $\mathcal{G} = \mathcal{G}(D)$ when we want to highlight its dependence on the policy D.) When the initial states satisfy $\hat{Q}^n(0) \Rightarrow \hat{Q}(0) \in \mathcal{G}$ (e.g., $\hat{Q}^n(0) = \hat{Q}(0) = 0$), the scaled routing processes defined in equation (18) satisfy the central limit theorem, i.e., the following weak convergence:

$$\hat{\Phi}^n(t) \Rightarrow \hat{\Phi}(t) := (\hat{\Phi}_k(t))_{k\in\mathcal{K}}, \tag{28}$$

where $\hat{\Phi}_k(t)$, $k \in \mathcal{K}$, are (possibly correlated) Brownian motions with zero means and finite and constant variations. Furthermore, putting the arrival, service and routing together, $(\hat{\Phi}(t), \hat{E}(t), \hat{S}(t))$ also constitutes a multi-dimensional Brownian motion with a zero drift and a finite constant covariance matrix.

We will see that an admissible policy indeed induces a well-defined diffusion limit for the queue length process under heavy traffic. Denote the class of all admissible routing policies as $\mathcal{D}$.

We call p_k^n the (approximate) routing rate to server k in the n-th system. The (approximate) arrival rate to and traffic intensity of server k are then denoted as $p_k^n\lambda^n$ and $\rho_k^n := p_k^n\lambda^n/\mu_k$, respectively. Hence, the condition in (a) requires that the arrival rates to the queues $p_k^n\lambda^n$ are within the service capacities (rates) μ_k and approach the capacities proportionally. This condition also implies

$$p_k^n \to p_k := \frac{\mu_k}{\mu_{\mathcal{K}}}, \quad \sum_{k\in\mathcal{K}} \theta_k = \theta_{\mathcal{K}}. \tag{29}$$

Observe that the convergence in (28) holds if p_k^n is replaced by any $\tilde{p}_k^n$ satisfying $n(p_k^k - \tilde{p}_k^n) \to 0$, and therefore, the choice of p_k^n is not unique. Also note that should the routing sequences satisfy the law of large numbers too, i.e., for some constants $\tilde{p}_k^n \geq 0$,

$$\lim_{\ell\to\infty} \Phi_k^n(\ell)/\ell = \tilde{p}_k^n \quad \text{a.s.,} \quad \text{for all } n,$$

it can be seen that $n(\tilde{p}_k^n - p_k^n) \to 0$, which justifies p_k^n as an "approximate" rate. The condition (a) causes no loss of generality for our study since, if it does not hold, we can always focus on any convergent subsequence of $\{\theta^n\}$.

Unlike the parameter $\theta_{\mathcal{K}}$, which is given in (11) and is fixed, we have some room to adjust the parameters θ_k's when we try to find the optimal routing policy below. Also note from (29) that p_k is the limit of the sequence $\{p_k^n\}$ rather than the routing rate for a specific system. Though p_k takes on the same value as $\mu_k/\mu_{\mathcal{K}}$, it indicates intentionally that the value is arrived at via a sequence. This is similar to different roles of λ and $\mu_{\mathcal{K}}$, and is useful for constructing routing policy for the original (pre-limit) system from the diffusion limit theorems (to be established) heuristically.

The condition (b) imposes the central limit theorem on the routing processes $\hat{\Phi}^n(t)$. There is vast amount of literature for characterizing conditions for the central limit theorem to hold, which relate to the notion of asymptotic independence. For examples, refer to: Gut [27], Section

5 of Chapter 9; Whitt [58], Section 4.4; and Shiryaev [50], Section 8 of Chapter VII. From a practical perspective, a useful routing policy should lead to an asymptotically stationary system, or simply a stationary system if it starts from a "nice" initial state, and the resulting routings, $\{\phi^n(\ell), \ell = 1, 2, \cdots\}$, must be an asymptotically independent sequence. Therefore, we believe that the class $\mathcal{D}$ contains a wide range of non-anticipating routing policies that are well behaved and of practical interest, which indeed includes all the policies being studied. Particularly, it includes some policies, such as the JSQ and BR policies (cf. Proposition 2), that are asymptotically optimal over the set of all routing policies (a strict superset of $\mathcal{D}$). By restricting our attention to the more tractable class $\mathcal{D}$, we are able to identify the optimal policies and characterize the system performances given the availability of state information such as the queue length state, the arrival history, and the service history.

3.1 Diffusion Limit and Lower Bound

The proposition below illustrates that an admissible policy induces a diffusion limit, which is amenable to further analysis, and establishes a lower bound for the class of admissible policies. This lower bound is indeed achievable by applying the JSQ and BR policies, as shown in Proposition 2.

Proposition 1 Suppose the routing policy D is admissible ($D \in \mathcal{D}$), and the initial states satisfy

$$\hat{Q}^n(0) \Rightarrow \hat{Q}(0) \in \mathcal{G}(D). \tag{30}$$

(a) (Diffusion limit) We have the following weak convergence: as $n \to \infty$,

$$\left(\hat{Q}^n(t), \hat{X}^n(t), \hat{Y}^n(t))\right) \Rightarrow \left(\hat{Q}(t), \hat{X}(t), \hat{Y}(t)\right), \tag{31}$$

where the limit is the unique solution of the following Skorohod problem:

$$\hat{Q}_k(t) = \hat{Q}_k(0) + \hat{X}_k(t) + \hat{Y}_k(t) \geq 0, \tag{32}$$

$$\hat{Y}_k(t) \text{ is non-decreasing in } t \text{ with } \hat{Y}_k(0) = 0, \tag{33}$$

$$\int_0^\infty \hat{Q}_k(t) d\hat{Y}_k(t) = 0, \tag{34}$$

with $\hat{X}(t) = (\hat{X}_k(t))_{k \in \mathcal{K}}$,

$$\hat{X}_k(t) = \hat{\Phi}_k(\lambda t) + p_k \hat{E}(t) - \hat{S}_k(t) + \theta_k t. \tag{35}$$

(b) (Lower bound) Moreover, we have the following asymptotic lower bound for the system:

$$(\hat{Q}_{\mathcal{K}}(t), \hat{Y}_{\mathcal{K}}(t)) \geq (\hat{Q}^*(t), \hat{Y}^*(t)), \tag{36}$$

where $(\hat{Q}^*(t), \hat{Y}^*(t))$ is the unique solution to the following Skorohod problem,

$$\hat{Q}^*(t) = \hat{Q}_{\mathcal{K}}(0) + \hat{X}_{\mathcal{K}}(t) + \hat{Y}^*(t) \geq 0, \tag{37}$$

$$\hat{Y}^*(t) \text{ is non-decreasing in } t \text{ with } \hat{Y}^*(0) = 0, \tag{38}$$

$$\int_0^\infty \hat{Q}^*(t) d\hat{Y}^*(t) = 0, \tag{39}$$

and

$$\hat{X}_{\mathcal{K}}(t) \left(= \sum_{k \in \mathcal{K}} \hat{X}_k(t) \right) = \hat{E}(t) - \sum_{k \in \mathcal{K}} \hat{S}_k(t) + \theta_{\mathcal{K}} t.$$

The proposition is proved by applying the continuity and the minimality properties of the one-dimensional Skorohod mapping (Lemma 10); a detailed proof is given in the Appendix.

It is known that given $\hat{Q}_k(0)$ and $\hat{X}_k(t)$, the relations in (32-34), which constitute the Skorohod problem, uniquely define the processes $\hat{Q}_k$ and $\hat{Y}_k$: $\hat{Q}_k = \Phi(\hat{Q}_k(0) + \hat{X}_k)$ and $\hat{Y}_k = \Psi(\hat{Q}_k(0) + \hat{X}_k)$, with $\Phi(\cdot)$ and $\Psi(\cdot)$ being Lipschitz continuous mappings (cf. Lemma 10). In particular, when $\hat{X}_k(t)$ is a Brownian motion, $\hat{Q}_k(t)$ is a one-dimensional *reflected* Brownian motion (RBM), and $\hat{Y}_k(t)$ is the associated *regulator*. As the free process $\hat{X}_k(t)$ has a negative drift $\theta_k < 0$, the process $\hat{Q}_k(t)$ has a stationary distribution, which is exponential with rate $-2\theta_k/\mathsf{var}(\hat{\Phi}_k(\lambda) + p_k\hat{E}(1) - \hat{S}_k(1))$ (cf. Lemma 11). Therefore, under the policy D (specified in the above proposition) the expected stationary queue length in the limit is

$$\mathsf{E}\hat{Q}_k(\infty; D) = \frac{\mathsf{var}(\hat{\Phi}_k(\lambda) + p_k\hat{E}(1) - \hat{S}_k(1))}{-2\theta_k}. \tag{40}$$

Here, $\hat{Q}_k(\infty; D)$, $k \in \mathcal{K}$, is a random variable following the stationary distribution of $\hat{Q}_k(t)$. In this notation, the routing policy D is attached to *explicitly* indicate that the variable (as well as the associated diffusion limit $\hat{Q}_k(t)$) is derived under the policy D, and such an argument will be omitted if it causes no confusion.

Similarly, the lower bound described in (37-39) can be characterized by $\hat{Q}^* = \Phi(\hat{Q}_{\mathcal{K}}(0)+\hat{X}_{\mathcal{K}})$ and $\hat{Y}^* = \Psi(\hat{Q}_{\mathcal{K}}(0) + \hat{X}_{\mathcal{K}})$, where the free process $\hat{X}_{\mathcal{K}}(t)$ has a negative drift $\theta_{\mathcal{K}}$ and a variation $\lambda c_a^2 + \sum_{k\in\mathcal{K}} \mu_k c_{b,k}^2$. The expected stationary queue length in the limit is

$$\mathsf{E}\hat{Q}^*(\infty) = \frac{\lambda c_a^2 + \sum_{k\in\mathcal{K}} \mu_k c_{b,k}^2}{-2\theta_{\mathcal{K}}} := \hat{L}^*, \tag{41}$$

where $\hat{Q}^*(\infty)$ follows the stationary distribution of $\hat{Q}^*(t)$. Clearly, it is the lower bound performance over all admissible policy, i.e.,

$$\mathsf{E}\hat{Q}^*(\infty) \le \mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; D), \quad D \in \mathcal{D}.$$

Furthermore, it can be noted that $Q^*(t)$ is indeed a lower bound in a stronger sense of stochastic order and pathwise dominance (e.g., [16], Theorem 5(b) and its proof) over all feasible routing policies, which might even not be non-anticipating or induce a diffusion limit. Nevertheless, it is sufficient for our purpose to restrict our discussion to class $\mathcal{D}$.

3.2 JSQ and BR Policies Revisited

Proposition 2 Consider the JSQ policy and the BR policy (with $c \ge 2$ and $\pi_k^n = \mu_k/\mu_{\mathcal{K}}$). (a) ([16, 47, 68]) Suppose the initial states satisfy

$$\hat{Q}^n(0) \Rightarrow \hat{Q}(0) \in \mathcal{G}(JSQ) := \{q : q \in \mathbb{R}_+^K,\ q_1 = \cdots = q_K\};$$

that is, all queue lengths are asymptotically equal. Then, under either the JSQ or BR policy, we have the weak convergence: as $n \to \infty$,

$$\left(\hat{Q}^n(t), \hat{X}^n(t), \hat{Y}^n(t))\right) \Rightarrow \left(\hat{Q}(t), \hat{X}(t), \hat{Y}(t)\right), \tag{42}$$

where in the diffusion limit the lower bound performance is achieved:

$$(\hat{Q}_{\mathcal{K}}(t), \hat{X}_{\mathcal{K}}(t), \hat{Y}_{\mathcal{K}}(t)) = (\hat{Q}^*(t), \hat{X}_{\mathcal{K}}(t), \hat{Y}^*(t)),$$

and all queues are equal: for all $k \in \mathcal{K}$,

$$(\hat{Q}_k(t), \hat{X}_k(t), \hat{Y}_k(t)) = \frac{1}{K}(\hat{Q}^*(t), \hat{X}_{\mathcal{K}}(t), \hat{Y}^*(t)), \tag{43}$$

$$\hat{\Phi}_k(\lambda t) = \left(\frac{1}{K} - p_k\right)\hat{E}(t) - \left(\frac{1}{K}\sum_{j\in\mathcal{K}} \hat{S}_j(t) - \hat{S}_k(t)\right), \quad \text{and} \tag{44}$$

$$\theta_k = \theta_{\mathcal{K}}/K. \tag{45}$$

(b) The JSQ policy and BR policy belong to the class $\mathcal{D}$, where the approximate routing rates and traffic intensities, p_k^n and ρ_k^n, can be chosen such that

$$\mu_k - p_k^n \lambda^n = \frac{1}{K}(\mu_{\mathcal{K}} - \lambda^n), \quad 1 - \rho_k^n = \frac{\mu_{\mathcal{K}}}{K\mu_k}(1 - \rho^n). \tag{46}$$

Observe that under the JSQ (or BR) policy, the diffusion limit attains the lower bound system given in Proposition 1(b), i.e., a one-dimensional RBM with drift $\theta_{\mathcal{K}}$ and variation $\lambda c_a^2 + \sum_k \mu_k c_{b,k}^2$. Hence, the expected stationary queue lengths for the whole system and each server are (cf. (40)),

$$\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; JSQ) = \frac{\lambda c_a^2 + \sum_k \mu_k c_{b,k}^2}{-2\theta_{\mathcal{K}}} \ (= \hat{L}^*), \tag{47}$$

$$\mathsf{E}\hat{Q}_k(\infty; JSQ) = \frac{1}{K}\hat{L}^*. \tag{48}$$

From equality (43), we see that the *state-space collapse* property holds, that is, the queue length of the diffusion limit evolves within the one-dimensional space $\mathcal{G}(JSQ)$. This subspace is called the fixed-point state space or invariant manifold in literature. Furthermore, the total queue length in the diffusion limit, $\hat{Q}_{\mathcal{K}}(t)$, is same as the one for a sequence of $G/G/1$ queueing systems — in the n-th system, the interarrival arrival times have a mean $1/\lambda^n$ and a squared coefficient of variation $(c_a^n)^2$, and the service times have a mean $1/\mu_{\mathcal{K}}$ and a squared coefficient of variations of $\sum_{k\in\mathcal{K}} \mu_k c_{b,k}^2/\mu_{\mathcal{K}}$. That is, the (original) parallel server system behaves like a $G/G/1$ queue and its servers appear pooled together to form an aggregated server in the diffusion limit. This is known as the *resource pooling* effect, which also exhibited for some other stochastic network models in previous studies (e.g., [32, 37, 43]).

Moreover, the initial states are required to "collapse" into the subspace $\mathcal{G}(JSQ)$ with all queue lengths being equal in the proposition. If we relax this requirement and allow initial states to be a tight sequence, then the initial states will jump to the fixed-point state space instantaneously. Consequently, the weak convergence in (42) still holds with a modification on the convergence of the inital period (cf. [9, 43, 52, 66]). This extension is indeed required when we study the interchange of limits below (cf. Propositions 20 and 21).

4 Round-Robin Routing: No State Information

Consider a sub-class of $\mathcal{D}$, denoted $\mathcal{H}$, which includes all admissible policies such that $\hat{\Phi}(\lambda t)$ and $(\hat{E}(t),\ \hat{S}(t))$, the limits derived in Proposition 1, are not correlated for all $t \geq 0$. Let $\mathcal{H}_0$ be a further sub-class of $\mathcal{H}$ containing the blind routing policies such that the routing $\{\Phi^n(\ell), \ell = 1, 2, \cdots\}$ is independent of $\{(E^n(t), S^n(t)), t \geq 0\}$. Hence, under any policy in $\mathcal{H}$ the routing is uncorrelated with the arrival and service history asymptotically, whereas under any policy in $\mathcal{H}_0$ the routing is independent of the arrival and service history. Also observe that the system dynamics, mainly described by $Q^n(t)$, is driven by the primitive processes $(E^n(t), S^n(t))$ and the routing control $\Phi^n(\ell)$ (cf. (22-24)). Thus, by using a blind policy (in $\mathcal{H}_0$), the controller can "see" at most its own decisions $\{\Phi^n(\ell)\}$. In contrast, the JSQ policy will review the queue length state, which has encoded the information about the arrival, service and routing available up to the epoch for routing decision-making.

For any policy $H \in \mathcal{H}$, since it belongs to $\mathcal{D}$ as well, Proposition 1 applies, and following the equality in (40) the expected stationary queue lengths of the diffusion limit are written as:

$$\mathsf{E}\hat{Q}_k(\infty; H) = \frac{\mathsf{var}(\hat{\Phi}_k(\lambda)) + p_k^2 \lambda c_a^2 + \mu_k c_{b,k}^2}{-2\theta_k}.$$

Observe that the expected stationary queue length is lower bounded by the following, given the same drift parameter $\theta(<0)$:

$$\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty;H) \geq h(\theta) := \sum_{k\in\mathcal{K}} \frac{p_k^2\lambda c_a^2 + \mu_k c_{b,k}^2}{-2\theta_k}.$$

Solving the optimization problem,

$$\min_{\theta_k}\ h(\theta),\quad s.t.\ \sum_{k\in\mathcal{K}}\theta_k = \theta_{\mathcal{K}},\ \theta_k < 0 \text{ for } k\in\mathcal{K}, \tag{49}$$

yields a lower bound of expected stationary queue lengths over all class $\mathcal{H}$ policies,

$$h(\theta^*) = \frac{\left(\sum_k \sqrt{p_k^2\lambda c_a^2 + \mu_k c_{b,k}^2}\right)^2}{-2\theta_{\mathcal{K}}},\quad \text{with } \theta^* = (\theta_k^*)_{k\in\mathcal{K}},\ \theta_k^* = \frac{\sqrt{p_k^2\lambda c_a^2 + \mu_k c_{b,k}^2}}{\sum_j \sqrt{p_j^2\lambda c_a^2 + \mu_j c_{b,j}^2}}\theta_{\mathcal{K}}. \tag{50}$$

Now, for the n-the system, apply the RR control policy (as described in Section 2) with the weight parameter $p^n = (p_k^n)_{k\in\mathcal{K}}$ satisfying the requirement in (2). We further require that the condition in (27) is satisfied for some $\theta = (\theta_k)_{k\in\mathcal{K}} < 0$, and denote such an RR policy sequence as $RR(\{p^n\}_{n\in\mathcal{N}},\theta)$, or $RR(\theta)$ for short, since we will see that the parameters $\{p^n\}$ take effect on the limit only through θ. The following theorem shows that the routing policy $RR(\theta)$, a blind policy belonging to $\mathcal{H}_0$, can achieve the lower bound performance in (50) over the larger class $\mathcal{H}$.

Theorem 3 (a) Suppose the RR policy, $RR(\theta)$, is in force, and the initial states satisfy $\hat{Q}^n(0) \Rightarrow \hat{Q}(0) \in \mathcal{G}(RR(\theta)) := \mathbb{R}_+^K$. Then, the weak convergence described in (31-35) holds with $\hat{\Phi}_k(t)$ and $\hat{X}_k(t)$ satisfying

$$\hat{\Phi}_k(t) = 0,\quad \hat{X}_k(t) = p_k\hat{E}(t) - \hat{S}_k(t) + \theta_k t. \tag{51}$$

The "free process" $\hat{X}_k(t)$ is a Brownian motion with drift θ_k and variance $(p_k^2\lambda c_a^2 + \mu_k c_{b,k}^2)$. Hence, for each $k\in\mathcal{K}$, $\hat{Q}_k(t)$ is an RBM with drift θ_k and variation $p_k^2\lambda c_a^2 + \mu_k c_{b,k}^2$, and the expected stationary queue lengths for the server k is:

$$\mathsf{E}\hat{Q}_k(\infty;RR(\theta)) = \frac{p_k^2\lambda c_a^2 + \mu_k c_{b,k}^2}{-2\theta_k}. \tag{52}$$

(b) The routing policy $RR(\theta)$ belongs to $\mathcal{H}_0(\subset\mathcal{H}\subset\mathcal{D})$, where the approximate routing rates for the n-th system can be given as $p^n = (p_k^n)_{k\in\mathcal{K}}$.
(c) Let θ be set to $\theta^* = (\theta_k^*)_{k\in\mathcal{K}}$ given in (50), or alternatively, choose some approximate routing rates such that

$$\frac{\mu_k - p_k^n\lambda^n}{\mu_{\mathcal{K}} - \lambda^n} = \frac{\sqrt{(p_k^n)^2\lambda^n(c_a^n)^2 + \mu_k c_{b,k}^2}}{\sum_j \sqrt{(p_j^n)^2\lambda^n(c_a^n)^2 + \mu_j c_{b,j}^2}}. \tag{53}$$

Then, the RR policy $RR^* := RR(\theta^*)$ is asymptotically optimal in $\mathcal{H}$; that is, for any policy $H\in\mathcal{H}$, we have

$$\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty;RR^*) \leq \mathsf{E}\hat{Q}_{\mathcal{K}}(\infty;H).$$

Moreover, the expected stationary queue lengths of the diffusion limit under RR^* are given as,

$$\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty;RR^*) = \frac{\left(\sum_k \sqrt{p_k^2\lambda c_a^2 + \mu_k c_{b,k}^2}\right)^2}{-2\theta_{\mathcal{K}}}, \tag{54}$$

$$\mathsf{E}\hat{Q}_k(\infty;RR^*) = \frac{\sqrt{p_k^2\lambda c_a^2 + \mu_k c_{b,k}^2}\left(\sum_j \sqrt{p_j^2\lambda c_a^2 + \mu_j c_{b,j}^2}\right)}{-2\theta_{\mathcal{K}}}. \tag{55}$$

The key to proving the above theorem is to establish the convergence of the routing processes to zero, i.e., the first equality in (51). And this follows from the property that the RR policy dispatches jobs to each server in a deterministic manner, as stipulated by the requirement in (2). Consequently, similar to the proof of Proposition 2, we can apply the Skorohod mapping to establish the diffusion limit (process-wise convergence) and subsequently the other properties in the theorem.

Note that in part (c) of the above theorem, $(\mu_k - \lambda^n p_k^n)$ and $(\mu_{\mathcal{K}} - \lambda^n)$ are the surplus capacities of the server k and the whole system, respectively; and $((p_k^n)^2\lambda^n(c_a^n)^2 + \mu_k c_{b,k}^2)$ is the combined variability owing to the arrival and service processes of the class k. Hence, under the optimal RR policy, the overall surplus capacity is distributed to each server in proportional to the square root of its combined variability. This is reminiscent of the *square-root rule* in various queueing models (e.g., [7,12,62]). Furthermore, in the special case of Poisson arrival and exponential service ($c_a^n = 1$ and $c_{b,k} = 1$), the terms inside the square roots become $((p_k^n)^2\lambda^n + \mu_k)$, a combination of the arrival and service rates. Hence, our square-root rule appears to be a generalization of the conventional forms that involve the square root of either arrival rate or service rate.

4.1 Comparison and Observation

Comparison between the optimal RR policy and the JSQ/BR policy. As the JSQ and BR policies yield the same diffusion limit, we discuss JSQ only.

From the optimality of the JSQ policy in Proposition 2, we know that the expected stationary queue length under the optimal RR policy RR^* cannot be smaller than the JSQ policy. Here, we first compare their performances in (47,54) and reveal when an RR policy can achieve the performance under JSQ policy, i.e., the optimal performance over all admissible policies (class $\mathcal{D}$).

To this end, we focus on the case $c_a^2 > 0$, while the other degenerate case ($c_a^2 = 0$) involves a tedious discussion but yields little new insight. To compare the performances, consider the function:

$$\Delta(x_1, \cdots, x_K) := \frac{\left(\sum_k \sqrt{p_k^2\lambda c_a^2 + x_k}\right)^2}{-2\theta_{\mathcal{K}}} - \frac{\lambda c_a^2 + \sum_k x_k}{-2\theta_{\mathcal{K}}}.$$

Clearly, we have

$$\Delta(0, \cdots, 0) = 0, \quad \Delta(\mu_1 c_{b,1}^2, \cdots, \mu_K c_{b,K}^2) = \mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; RR^*) - \mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; JSQ),$$

and when $(x_1, \cdots, x_K) \geq 0$,

$$\frac{\partial \Delta}{\partial x_k} = \frac{1}{-2\theta_{\mathcal{K}}} \frac{\sum_j \sqrt{p_j^2\lambda c_a^2 + x_j}}{\sqrt{p_k^2\lambda c_a^2 + x_k}} - \frac{1}{-2\theta_{\mathcal{K}}} = \frac{1}{-2\theta_{\mathcal{K}}} \sum_{j \in \mathcal{K}\setminus\{k\}} \frac{\sqrt{p_j^2\lambda c_a^2 + x_j}}{\sqrt{p_k^2\lambda c_a^2 + x_k}} > 0, \quad k \in \mathcal{K}.$$

This implies that the equality, $\Delta(\mu_1 c_{b,1}^2, \cdots, \mu_K c_{b,K}^2) = 0$, or $\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; RR^*) = \mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; JSQ)$, holds only when

$$c_{b,k}^2 = 0 \text{ for all } k \in \mathcal{K}.$$

Under the above condition, the parameters θ_k^* and p_k^n specified in Theorem 3(c) can be simplified as:

$$\theta_k^* = p_k\theta_{\mathcal{K}}, \quad p_k^n = \frac{\mu_k}{\mu_{\mathcal{K}}}.$$

In other words, the optimal RR policy can attain the performance of the JSQ policy if and only if all service times are deterministic, and in this case, jobs are routed to each server in proportional to the service rate.

Wu and Down [63] studied the same model under the RR routing, but assuming that the job arrivals follow a Poisson process and the service times follow a discrete requirement and are known upon arrival. Under the latter assumption, their model is reduced to the one with multiple job classes, each class having a Poisson arrival stream, deterministic service times, and priority scheduling at servers. Then, they derive the diffusion limit under the RR routing policy, and show (in Theorem 3.3) that the RR policy, assigning jobs in proportional to the service rates for each class, is asymptotically optimal and has the same diffusion limit as the JSQ policy, as well as the c-SRPT policy in their paper. Their optimality result (more specifically, in the special case with a single job class) is consistent with our observation above, i.e., the RR policy achieves the JSQ performance only when all service times are deterministic.

We now consider an example, and illustrate that the optimal RR policy can yield an expected stationary queue length arbitrarily longer than the JSQ policy. In this example, assume all servers are identical, services times are exponentials, and arrivals follow the Poisson process. Then, we have $p_k = 1/K$ and $c_a^2 = c_{b,k}^2 = 1$, and thus $\theta_k^* = \theta_{\mathcal{K}}/K$ according to the above discussion. Consequently, according to Theorems 2 and 3, the expected queue lengths are reduced to the following,

$$\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; RR^*) = \frac{(1+K)\lambda}{-2\theta_{\mathcal{K}}}, \quad \mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; JSQ) = \frac{\lambda}{-\theta_{\mathcal{K}}}.$$

When the number of servers K grows, the performance under the optimal RR policy can be arbitrarily worse than the JSQ policy, i.e.,

$$\frac{\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; RR^*)}{\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; JSQ)} \to \infty \quad \text{as } K \to \infty.$$

Comparison between the optimal RR policy and the proportional RR policy. For the RR policy $RR(\theta)$, a conventional option is, as we have seen above, to distribute the jobs to servers in proportion to the service rate, i.e., to set the parameters as $p_k^n = \mu_k/\mu_{\mathcal{K}}$ and thus from the condition in (27),

$$\theta_k = \theta'_k := \frac{\mu_k}{\mu_{\mathcal{K}}}\theta_{\mathcal{K}}. \tag{56}$$

We call it the proportional RR policy. From Theorem 3(a), the expected stationary queue length under this policy is

$$\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; RR(\theta')) = \sum_{k\in\mathcal{K}} \frac{p_k\lambda c_a^2 + \mu_{\mathcal{K}} c_{b,k}^2}{-2\theta_{\mathcal{K}}}. \tag{57}$$

Putting (57,54) together, and taking into account the relationships $p_k = \mu_k/\mu_{\mathcal{K}}$ and $\lambda = \mu_{\mathcal{K}}$, then yield the ratio between the performances under the proportional and the optimal RR policies:

$$\frac{\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; RR(\theta'))}{\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; RR(\theta^*))} = \frac{\sum_{k\in\mathcal{K}}(p_k c_a^2 + c_{b,k}^2)}{\left(\sum_{k\in\mathcal{K}}\sqrt{p_k(p_k c_a^2 + c_{b,k}^2)}\right)^2} \geq 1,$$

where the inequality is due to the conclusion in Theorem 3(c) and alternatively can be shown directly using the concavity of the square root function. Furthermore, let us examine an example: assume Poisson arrival and exponential service (hence, $c_a^2 = 1$, $c_{b,k}^2 = 1$), and pick $p_1 = 1 - 1/K + 1/K^2$ and $p_k = 1/K^2$ for $k = 2, \cdots, K$. Then, the above performance ratio is reduced to

$$\frac{\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; RR(\theta'))}{\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; RR(\theta^*))} = \frac{K+1}{3+2\sqrt{2}+o(1/K)} \to \infty, \quad \text{as } K \to \infty.$$

Therefore, the performance of the proportional RR policy is generally suboptimal within the class of RR policies, and can be arbitrarily worse than the optimal RR policy.

Next, observing the definitions of θ' and θ^* in (56) and (50), respectively, we note that the proportional RR policy coincides with the optimal RR policy if and only if

$$\frac{\sqrt{p_k^2\lambda c_a^2+\mu_k c_{b,k}^2}}{\sum_{j\in\mathcal{K}}\sqrt{p_j^2\lambda c_a^2+\mu_j c_{b,j}^2}}=p_k,\quad k\in\mathcal{K}. \tag{58}$$

By simple algebra, this is reduced to $\mu_1^{-1}c_{b,1}^2=\mu_2^{-1}c_{b,2}^2=\cdots=\mu_K^{-1}c_{b,K}^2$. Recall $\mu_k^{-1}c_{b,k}^2=\sigma_{b,k}^2/\mu_k^{-1}$, where $\sigma_{b,k}^2$ denotes the variation of the service time (of server k) and thus $\sigma_{b,k}^2/\mu_k^{-1}$ is the variance-to-mean ratio. Consequently, the proportional RR policy coincides with the optimal RR policy if and only if variance-to-mean ratios of service times for all servers are the same. Therefore, in the special case that all servers are identical, the (asymptotically) optimal RR policy will route incoming jobs in order and in rotating fashion. This policy is indeed the optimal routing policy for a discrete system with identical parallel servers, renewal arrival, general service time at each server, and no real-time information of the system state (cf. Hajek [28] and Altman *et al.* [1]).

5 Probabilistic Proportional Routing: No State Information

Consider the PP routing policy (as described in Section 2) with the weight parameter $p^n=(p_k^n)_{k\in\mathcal{K}}$ satisfying the requirement in (4) and the condition in (27) for some $\theta<0$. Denote such a PP policy sequence as $PP(\{p^n\}_{n\in\mathcal{N}},\theta)$, or $PP(\theta)$ for short. In the following theorem we identify the optimal PP policy.

Theorem 4 (a) Suppose the PP policy, $PP(\theta)$, is in force, and the initial states satisfy $\hat{Q}^n(0)\Rightarrow\hat{Q}(0)\in\mathcal{G}(PP(\theta)):=\mathbb{R}_+^K$. Then, the weak convergence depicted in (31-35) holds. In the limit, the routing process $\hat{\Phi}(\lambda t)=(\hat{\Phi}_k(\lambda t))_{k\in\mathcal{K}}$ is a K-dimensional Brownian motion that is independent of $(\hat{E}(t),\hat{S}(t))$ and has a correlation matrix $\Gamma=(\Gamma_{k\ell})_{\{k,\ell\in\mathcal{K}\}}$ with $\Gamma_{kk}=\lambda p_k(1-p_k)$ and $\Gamma_{k\ell}=-\lambda p_k p_\ell$ for $k\neq\ell$. The "free process" $\hat{X}_k(t)$ is a Brownian motion with drift θ_k and variance $(\lambda p_k(1-p_k)+p_k^2\lambda c_a^2+\mu_k c_{b,k}^2)$. Hence, $\hat{Q}_k(t)$ is an RBM with drift θ_k and variation $\lambda p_k(1-p_k)+p_k^2\lambda c_a^2+\mu_k c_{b,k}^2$for each $k\in\mathcal{K}$, and the expected stationary queue lengths are:

$$\mathsf{E}\hat{Q}_k(\infty;PP(\theta))=\frac{\lambda p_k(1-p_k)+p_k^2\lambda c_a^2+\mu_k c_{b,k}^2}{-2\theta_k}, \tag{59}$$

$$\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty;PP(\theta))=\sum_{k\in\mathcal{K}}\frac{\lambda p_k(1-p_k)+p_k^2\lambda c_a^2+\mu_k c_{b,k}^2}{-2\theta_k}. \tag{60}$$

(b) The routing policy $PP(\theta)$ belongs to $\mathcal{H}_0(\subset\mathcal{H}\subset\mathcal{D})$, where the approximate routing rates for the n-th system can be given as the weight parameter $p^n=(p_k^n)_{k\in\mathcal{K}}$.
(c) Let θ be set to $\theta^*=(\theta_k^*)_{k\in\mathcal{K}}$, with

$$\theta_k^*=\frac{\sqrt{\lambda p_k(1-p_k)+p_k^2\lambda c_a^2+\mu_k c_{b,k}^2}}{\sum_j\sqrt{\lambda p_j(1-p_j)+p_j^2\lambda c_a^2+\mu_j c_{b,j}^2}}\theta_{\mathcal{K}};$$

or alternatively, choose some approximate routing rates such that

$$\frac{\mu_k-p_k^n\lambda^n}{\mu_{\mathcal{K}}-\lambda^n}=\frac{\sqrt{\lambda^n p_k^n(1-p_k^n)+(p_k^n)^2\lambda^n(c_a^n)^2+\mu_k c_{b,k}^2}}{\sum_j\sqrt{\lambda^n p_j^n(1-p_j^n)+(p_j^n)^2\lambda^n(c_a^n)^2+\mu_j c_{b,j}^2}}. \tag{61}$$

Then, the policy $PP^* := PP(\theta^*)$ is asymptotically optimal over all PP policies; that is, for any PP policy $PP(\theta)$, we have

$$\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; PP^*) \leq \mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; PP(\theta)).$$

Moreover, the expected stationary queue lengths under the optimal PP policy PP^* are given as:

$$\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; PP^*) = \frac{\left(\sum_k \sqrt{\lambda p_k(1-p_k) + p_k^2\lambda c_a^2 + \mu_k c_{b,k}^2}\right)^2}{-2\theta_{\mathcal{K}}}, \tag{62}$$

$$\mathsf{E}\hat{Q}_k(\infty; PP^*) = \frac{\sqrt{\lambda p_k(1-p_k) + p_k^2\lambda c_a^2 + \mu_k c_{b,k}^2}\left(\sum_j \sqrt{\lambda p_j(1-p_j) + p_j^2\lambda c_a^2 + \mu_j c_{b,j}^2}\right)}{-2\theta_{\mathcal{K}}} \tag{63}$$

Conclusion (a) in the above theorem is a special case of the diffusion limit for the (generalized) Jackson network (e.g., [15]). Conclusion (b) is obvious, and (c) is established by choosing the best drift parameter θ to minimize the expected queue length given in (a). Detailed proof is omitted.

Similar to the optimal RR policy, the optimal PP policy, particularly the routing rate in (61), also exhibits a certain form of the square-root rule. In particular, in the case of Poisson arrival and exponential service (hence, $c_a^n = 1$ and $c_{b,k} = 1$) and for large n (hence, $\lambda^n p_k^n \approx \mu_k$), the routing rate is approximated as,

$$\mu_k - p_k^n\lambda^n \approx \frac{\sqrt{\mu_k}}{\sum_j \sqrt{\mu_j}}(\mu_{\mathcal{K}} - \lambda^n).$$

Therefore, in this special case, the overall surplus capacity $(\mu_{\mathcal{K}} - \lambda^n)$ is distributed to each server in proportion to the square root of the service rate μ_k.

Observe from the equalities in (52,60) that with the same drift parameter θ, the RR policy has an expected stationary queue length strictly less than the PP policy. Hence, the optimal RR policy also asymptotically outperforms the optimal PP policy strictly. Moreover, by comparing the equalities in (54,62), it is direct to see that the performance ratio between the optimal PP and RR policies, $\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; PP^*)/\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; RR^*)$, can be arbitrarily large when c_a and $c_{b,k}$ are sufficiently small.

6 Arrival Chasing: Using Arrival History Information

Consider a sub-class of $\mathcal{D}$, denoted $\mathcal{E}$, which includes all admissible policies such that $\hat{\Phi}(\lambda t)$ and $\hat{S}(t)$ are not correlated for $t \geq 0$. Let $\mathcal{E}_0$ be a further sub-class of $\mathcal{E}$ containing policies such that the routing $\{\Phi^n(\ell), \ell = 1, 2, \cdots\}$ is independent of $\{S^n(t), t \geq 0\}$ for each n.

Clearly, any policy $E \in \mathcal{E}$ belongs to $\mathcal{D}$ as well, and therefore we apply Proposition 1 (the equality in (40) in particular), along with the definition of $\mathcal{E}$, to write the expected stationary queue lengths as:

$$\mathsf{E}\hat{Q}_k(\infty; E) = \frac{\mathrm{var}(\hat{\Phi}_k(\lambda) + p_k\hat{E}(1)) + \mu_k c_{b,k}^2}{-2\theta_k}, \quad \mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; E) = \sum_{k\in\mathcal{K}} \mathsf{E}\hat{Q}_k(\infty; E). \tag{64}$$

A lower bound of $\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; E)$ can be estimated as follows. If we let

$$h_k = \frac{\mathsf{E}[\hat{\Phi}_k(\lambda)\hat{E}(1)]}{\mathsf{E}\hat{E}(1)^2},$$

which satisfies $\sum_{k\in\mathcal{K}} h_k = 0$ (since $\sum_{k\in\mathcal{K}} \hat{\Phi}_k(\lambda) = 0$); then, the variation terms in $\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; E)$ can be estimated as

$$\begin{aligned} & \mathsf{var}(\hat{\Phi}_k(\lambda) + p_k\hat{E}(1)) = \mathsf{var}\left((\hat{\Phi}_k(\lambda) - h_k\hat{E}(1)) + (h_k + p_k)\hat{E}(1)\right) \\ = \quad & \mathsf{var}\left((\hat{\Phi}_k(\lambda) - h_k\hat{E}(1))\right) + (h_k + p_k)^2\,\mathsf{var}(\hat{E}(1)) \geq (h_k + p_k)^2\lambda c_a^2, \end{aligned}$$

where the equality is attained if

$$\hat{\Phi}_k(\lambda t) = h_k\hat{E}(t). \tag{65}$$

Hence, we have the following for some constants $h = (h_k)_{k\in\mathcal{K}}$ satisfying $\sum_{k\in\mathcal{K}} h_k = 0$,

$$\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; E) \geq f(\theta, h) := \sum_{k\in\mathcal{K}} \frac{(h_k + p_k)^2\lambda c_a^2 + \mu_k c_{b,k}^2}{-2\theta_k}.$$

And consequently, solving the following optimization problem will yield a lower bound for $\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; E)$ over the class $\mathcal{E}$:

$$\min_{\theta_k, h_k} \; f(\theta, h), \quad s.t. \; \sum_{k\in\mathcal{K}} h_k = 0, \; \sum_{k\in\mathcal{K}} \theta_k = \theta_{\mathcal{K}}. \tag{66}$$

Applying the KKT condition, it is direct to solve the above for the optimal solution (h_k^*, θ_k^*)

$$h_k^* + p_k = \frac{\theta_k^*}{\theta_{\mathcal{K}}} = \frac{\sqrt{\mu_k c_{b,k}^2}}{\sum_j \sqrt{\mu_j c_{b,j}^2}}, \tag{67}$$

and the lower bound for $\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; E)$,

$$f(\theta^*, h^*) = \frac{\lambda c_a^2 + \left(\sum_k \sqrt{\mu_k c_{b,k}^2}\right)^2}{-2\theta_{\mathcal{K}}}, \tag{68}$$

with the corresponding estimate for $\mathsf{E}\hat{Q}_k(\infty; E)$:

$$f_k(\theta^*, h^*) = (h_k^* + p_k) f(\theta^*, h^*). \tag{69}$$

(Here and below, some constants and parameters (e.g., θ^* and h^*) are redefined to represent different values. As this appears under different classes of routing policies, we hope no confusion will arise.)

Next, we introduce the so-called *arrival-chasing* (AC) routing policy in the class $\mathcal{E}_0$, which will indeed lead to a limit in the form of (65) and achieve the lower bound in (68).

6.1 Arrival-Chasing Policy and Its Optimality

An AC routing policy for the n-th network is specified with the parameters $p^n = (p_k^n)_{k\in\mathcal{K}} \geq 0$ and $h^n = (h_k^n)_{k\in\mathcal{K}}$ satisfying

$$\sum_{k\in\mathcal{K}} p_k^n = 1 \quad \text{and} \quad \sum_{k\in\mathcal{K}} h_k^n = 0 \tag{70}$$

as follows: Upon the ℓ-th arrival ($\ell = 1, 2, \cdots$), pick any server k' such that

$$k' = \arg\min_{k\in\mathcal{K}} \alpha_k^n(\ell), \quad \text{with } \alpha_k^n(\ell) := (\Phi_k^n(\ell - 1) - p_k^n\ell) - h_k^n(\ell - \lambda^n\Upsilon^n(\ell)), \tag{71}$$

and then let

$$\Phi^n_{k'}(\ell) = \Phi^n_{k'}(\ell-1) + 1 \quad \text{and} \quad \Phi^n_k(\ell) = \Phi^n_k(\ell-1),\ k \neq k'. \tag{72}$$

If we denote $t = \Upsilon^n(\ell)$ and $\tau = \Upsilon^n(\ell-1)$, then, the function $\alpha^n_k(\ell)$, defined for choosing server in the above, can be written as:

$$\alpha^n_k(\ell) = (\Phi^n_k(E^n(\tau)) - p^n_k E^n(t)) - h^n_k(E^n(t) - \lambda^n t). \tag{73}$$

Take a closer look into how a server is selected upon the arrival of a job in (73) (or (71)). Note that the term $E^n(t) - \lambda^n t$, called "arrival deviation" here, is the (real-time) deviation of the external arrival from its mean. A portion of this arrival deviation will be assigned to server k as specified by the parameter h^n_k (which could be either positive or negative), and acts as the "targeted deviation". Given the parameter p^n_k—the long-run fraction of jobs assigned to server k, the term $p^n_k E^n(t)$ can be viewed as the nominal amount of jobs dispatched to server k, or the "mean routing". Then, the term $\Phi^n_k(E^n(\tau)) - p^n_k E^n(t)$ is the (real-time) deviation from the mean routing, and can be called "routing deviation". Now, the difference, $\alpha^n_k(\ell)$, represents the "surplus" of routing deviation to be minimized. Thus, the AC policy steers the routing deviation to "chase" the targeted deviation — hence, the name arrival-chasing. Observe that when $h^n = 0$ (i.e., the history of past arrivals is ignored), the AC policy will be reduced to an RR policy.

It would be useful to define a chasing process, $\Psi^n(t) = (\Psi^n_k(t))_{k\in\mathcal{K}}$:

$$\Psi^n_k(t) := (\Phi^n_k(E^n(t)) - p^n_k E^n(t))) - h^n_k(E^n(t) - \lambda^n t). \tag{74}$$

Note that $\alpha^n_k(\ell)$ gives some state information (upon the ℓ-th arrival at time t) excluding the routing information of the arriving job, while the process $\Psi^n(t)$ has embodied the routing information of that job. Clearly, at time $t = \Upsilon^n(\ell)$, we have

$$\Psi^n_k(t) = \alpha^n_k(\ell) + \phi^n_k(\ell), \quad \sum_{k\in\mathcal{K}} \Psi^n_k(t) = \sum_{k\in\mathcal{K}} \alpha^n_k(t) + 1 = 0. \tag{75}$$

The AC policy can be implemented iteratively. Given $\Psi^n_k(\tau)$, the information of the chasing process upon the $(\ell-1)$-th arrival (or, time $\tau = \Upsilon^n(\ell-1)$), it is only necessary to monitor the interarrival time $u^n(\ell)$ till the next arrival, and then, calculate $\alpha^n_k(\ell)$ as

$$\alpha^n_k(\ell) = \Psi^n_k(\tau) - p^n_k - h^n_k(1 - \lambda^n u^n(\ell)). \tag{76}$$

This determines the routing $\phi^n(\ell)$ (or $\Phi^n(\ell)$) following (71,72), and thus the chasing process at time $t = \Upsilon^n(\ell)$ (of the ℓ-th arrival) is updated as in (75). Clearly, given the parameters $(p^n_k, h^n_k, \lambda^n)$, the routing decision of an AC policy depends only on the information about the arrival and routing available up to the decision epoch, and is independent of the service process. And, as we will see in the next theorem, to specify the control parameters (p^n_k, h^n_k) optimally requires a priori knowledge of the model parameters $(\lambda^n, \mu_k, c_{b,k})$.

Now, apply the AC routing policy defined in (71,72) with the parameters (p^n, h^n) satisfying the requirements in (70). We further require that the condition in (27) is satisfied for some $\theta < 0$, and

$$h^n_k \to h_k, \quad \text{as } n \to \infty, \text{ for } k \in \mathcal{K}, \tag{77}$$

for some $h = (h_k)_{k\in\mathcal{K}}$ satisfying $\sum_{k\in\mathcal{K}} h_k = 0$. Denote such an AC policy sequence as $AC(\{p^n, h^n\}_{n\in\mathcal{N}}, \theta, h)$, or $AC(\theta, h)$ for short. The following theorem shows that the routing policy $AC(\theta, h)$ belongs to $\mathcal{E}_0$, and can achieve the lower bound performance in (68) over a larger class $\mathcal{E}$.

Theorem 5 (a) Suppose the AC policy, $AC(\theta, h)$, is in force, and the initial states satisfy $\hat{Q}^n(0) \Rightarrow \hat{Q}(0) \in \mathcal{G}(AC(\theta,h)) := \mathbb{R}_+^K$. Then, the weak convergence depicted in (31-35) holds with $\hat{\Phi}_k(t)$ satisfying (65) and

$$\hat{X}_k(t) = (h_k + p_k)\hat{E}(t) - \hat{S}_k(t) + \theta_k t.$$

Hence, $\hat{Q}_k(t)$ is an RBM with drift θ_k and variation $(h_k+p_k)^2\lambda c_a^2 + \mu_k c_{b,k}^2$, and the expected stationary queue lengths are:

$$\mathsf{E}\hat{Q}_k(\infty; AC(\theta,h)) = \frac{(h_k+p_k)^2\lambda c_a^2 + \mu_k c_{b,k}^2}{-2\theta_k}, \tag{78}$$

$$\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; AC(\theta,h)) = \sum_{k\in\mathcal{K}} \frac{(h_k+p_k)^2\lambda c_a^2 + \mu_k c_{b,k}^2}{-2\theta_k}. \tag{79}$$

(b) The routing policy $AC(\theta,h)$ belongs to $\mathcal{E}_0 (\subset \mathcal{E} \subset \mathcal{D})$, where the approximate routing rates for the n-th system can be given as $p^n = (p_k^n)_{k\in\mathcal{K}}$.
(c) Let (θ, h) be set to (θ^*, h^*) given in (67), or alternatively, choose the sequence of parameters $\{p^n, h^n\}_{n\in\mathcal{N}}$ as

$$h_k^n + p_k^n = \frac{\mu_k - p_k^n\lambda^n}{\mu_{\mathcal{K}} - \lambda^n} = \frac{\sqrt{\mu_k c_{b,k}^2}}{\sum_j \sqrt{\mu_j c_{b,j}^2}}. \tag{80}$$

Then, the policy $AC^* \equiv AC(\theta^*, h^*)$ is asymptotically optimal in $\mathcal{E}$; that is, for any policy $E \in \mathcal{E}$, we have

$$\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; AC^*) \le \mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; E).$$

Moreover, the expected stationary queue lengths under AC^* are given in (68,69), i.e.,

$$\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; AC^*) = \frac{\lambda c_a^2 + \left(\sum_j \sqrt{\mu_j c_{b,j}^2}\right)^2}{-2\theta_{\mathcal{K}}}, \tag{81}$$

$$\mathsf{E}\hat{Q}_k(\infty; AC^*) = (h_k^* + p_k)\cdot \mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; AC^*) = \frac{\sqrt{\mu_k c_{b,k}^2}}{\sum_j \sqrt{\mu_j c_{b,j}^2}} \mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; AC^*). \tag{82}$$

Similar to the proof of Theorem 3, the key is to show the convergence of the routing processes to the limit in (65). Under the AC policy, the routing deviation chases the arrival deviation (adjusted with the coefficient h_k^n); refer to (71) or (73). With this property, we can apply the hydrodynamic approach to establish the required convergence (cf. Lemmas 14 and 15).

From the above theorem, we observe that a square-root rule exhibits in the optimal AC policy once again, similar to the optimal RR and PP policies. Completely analogous to the optimal PP policy, in the case of Poisson arrival and exponential service (hence, $c_a^n = 1$ and $c_{b,k} = 1$) the routing rate under the optimal SC policy, given in equation (80), is written as,

$$\mu_k - p_k^n\lambda^n = \frac{\sqrt{\mu_k}}{\sum_j \sqrt{\mu_j}}(\mu_{\mathcal{K}} - \lambda^n).$$

That is, the overall surplus capacity is distributed to each server in proportion to the square root of the service rate.

A subtle point relates to the trivial case that all servers are deterministic ($c_{b,k} = 0$ for all k). In this case, it can be checked that part (c) of the theorem still holds if the coefficient in (80,82) (or, the last term in (67)), which has a zero in the denominator, is interpreted as an arbitrarily valid weight parameter $w = (w_k)_{k\in\mathcal{K}}$ satisfying $w \ge 0$ and $\sum_{k\in\mathcal{K}} w_k = 1$. In other words, it only requires the first equality in (67) for the AC policy to be optimal. We omit a detailed discussion of this case to avoid complicating the presentation.

6.2 Comparison and Observation

Comparison between the optimal AC policy and the JSQ policy. Since the JSQ policy asymptically minimizes the expected stationary queue length globally (even beyond class $\mathcal{D}$), it performs better than the optimal AC policy, i.e.,

$$\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; AC^*) \geq \mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; JSQ) \ (= \hat{L}^*).$$

From the expressions in (81,47), we note that the equality takes effect if and only if $c_{b,k} > 0$ for at most one of the servers. That is, when all servers, except at most one, have deterministic service times, the optimal AC policy AC^*, using only the information of arrival history, can achieve the JSQ performance.

In addition, if the service time for some server k is deterministic ($c_{b,k} = 0$), we have $h_k^* + p_k = \theta_k^* = 0$ from (67). Thus, using the optimal AC policy, we will choose $h_k^n = -p_k^n$ and $p_k^n \lambda^n = \mu_k$ for the n-th system, according to Theorem 5(c). Recall the definition of the AC policy in (73), and note that the job stream routed to a server k is generally variable inherent in the use of arrival information,

$$\Phi_k^n(E^n(t)) \approx (p_k^n + h_k^n)E^n(t) - h_k^n \lambda^n t.$$

Then, given the above choice of parameters, the arrival stream to server k becomes (nearly) deterministic,

$$\Phi_k^n(E^n(t)) \approx p_k^n \lambda^n t = \mu_k t.$$

Therefore, the optimal AC policy strikes to cancel out the variability of the stream routed to any deterministic server. Given the deterministic arrival and service processes, such a server can be fully utilized while its queue length maintains near zero.

Comparison between the optimal AC policy and the optimal RR policy. Recall that the optimal RR and AA policies, described in Theorems 3(c) and 5(c) respectively, are specified by the optimal solutions to the problems in (49) and (66) respectively. Both problems have a strictly convex objective function and thus have their unique solutions (cf. (50) and (67)). Comparing these two problems yields that

$$\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; AC^*) \leq \mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; RR^*),$$

and that the equality in the above holds if and only if $h_k^* = 0$ for all k. From the equalities in (67), it is direct to verify that the latter condition holds if and only if $\mu_k^{-1} c_{b,k}^2$, $k \in \mathcal{K}$, are all equal. Recall that $\mu_k^{-1} c_{b,k}^2 = \sigma_{b,k}^2 / \mu_k^{-1}$, where $\sigma_{b,k}^2$ denotes the variance of class-k service times. In other words, when the variance-to-mean ratio of service times at all servers are equal, both the optimal RR and AC policies yields the same expected stationary queue length, and thus using the arrival information does not help minimize total queue length.

Next, we investigate to what extent the arrival information can help improve performance — the value of arrival information.

For the performances under the optimal RR and AC policies in (54,81), we have

$$\begin{aligned} &\left(\sum_{k \in \mathcal{K}} \sqrt{p_k^2 \lambda c_a^2 + \mu_k c_{b,k}^2} \right)^2 \leq \left(\sum_{k \in \mathcal{K}} \left(\sqrt{p_k^2 \lambda c_a^2} + \sqrt{\mu_k c_{b,k}^2} \right) \right)^2 \\ = \ &\left(\sqrt{\lambda c_a^2} + \sum_{k \in \mathcal{K}} \sqrt{\mu_k c_{b,k}^2} \right)^2 \leq 2 \left(\lambda c_a^2 + \left(\sum_{k \in \mathcal{K}} \sqrt{\mu_k c_{b,k}^2} \right)^2 \right), \end{aligned}$$

where the first inequality is due to the triangle inequality and the second to the convexity of square function. Note that there must be a strict inequality in the above. Otherwise, the equality in the first inequality would require that either $p_k^2 \lambda c_a^2$ or $\mu_k c_{b,k}^2$ must be zero for each k, i.e.,

$c_a^2 = 0$ or $(c_{b,k}^2)_{k\in\mathcal{K}} = 0$. This would prohibit the equality in the second inequality. (Recall, the trivial case that c_a and $c_{b,k}$ are all zero is not considered in this paper.) Hence, the performance ratio between two policies is strictly less than 2:

$$\frac{\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; RR^*)}{\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; AC^*)} = \frac{\left(\sum_k \sqrt{p_k^2\lambda c_a^2 + \mu_k c_{b,k}^2}\right)^2}{\lambda c_a^2 + \left(\sum_k \sqrt{\mu_k c_{b,k}^2}\right)^2} < 2.$$

Moreover, if we pick some parameters such that $c_a^2 > 0$, $c_{b,k}^2 = 0$ for $k = 2, \cdots, K$, and $c_{b,1}^2 = (1/p_1 + 2)c_a^2$, by simple calculation, we can reduce the above ratio to:

$$\frac{\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; RR^*)}{\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; AC^*)} = \frac{2}{1+p_1}.$$

Therefore, the ratio can be arbitrarily close to 2, the upper bound ratio, if p_1 is sufficiently small.

In summary, as compared with the optimal RR policy, the optimal AC policy can reduce up to 50% of the expected stationary queue length by utilizing the arrival information.

7 Service Chasing: Using Service History Information

Consider a sub-class of $\mathcal{D}$, denoted $\mathcal{S}$, which includes all admissible policies such that $\hat{\Phi}(\lambda t)$ and $\hat{E}(t)$ are not correlated for $t \geq 0$.

Since $\mathcal{S} \subset \mathcal{D}$, Proposition 1 applies to any routing policy in $\mathcal{S}$. Hence, under any policy $S \in \mathcal{S}$, the expected stationary queue lengths can be written as

$$\mathsf{E}\hat{Q}_k(\infty; S) = \frac{p_k^2\lambda c_a^2 + \mathrm{var}(\hat{\Phi}_k(\lambda) - \hat{S}_k(1))}{-2\theta_k}. \tag{83}$$

Below we derive a lower bound for $\mathsf{E}\hat{Q}_{k\in\mathcal{K}}(\infty; S)$ in a way similar to the AC policy in Section 6. Denote $h = (h_{ki})_{k,i\in\mathcal{K}}$, with

$$\begin{aligned} 1 - h_{kk} &= \frac{\mathsf{E}[\hat{\Phi}_k(\lambda)\hat{S}_k(1)]}{\mathsf{E}\hat{S}_k(1)^2}, \quad \text{for } k \in \mathcal{K}, \\ -h_{ki} &= \frac{\mathsf{E}[\hat{\Phi}_k(\lambda)\hat{S}_i(1)]}{\mathsf{E}\hat{S}_i(1)^2}, \quad \text{for } k \neq i,\ k, i \in \mathcal{K}, \end{aligned}$$

which satisfies for all $i \in \mathcal{K}$,

$$1 - \sum_{k\in\mathcal{K}} h_{ki} = \frac{\mathsf{E}[\sum_{k\in\mathcal{K}} \hat{\Phi}_k(\lambda)\hat{S}_i(1)]}{\mathsf{E}\hat{S}_i(1)^2} = 0.$$

The variation term in $\mathsf{E}\hat{Q}_k(\infty; S)$ can be estimated as

$$\begin{aligned} &\mathsf{var}(\hat{\Phi}_k(\lambda) - \hat{S}_k(1)) = \mathrm{var}\left(\left(\hat{\Phi}_k(\lambda) - \left(\hat{S}_k(1) - \sum_{i\in\mathcal{K}} h_{ki}\hat{S}_i(1)\right)\right) - \sum_{i\in\mathcal{K}} h_{ki}\hat{S}_i(1)\right) \\ = \ &\mathsf{var}\left(\hat{\Phi}_k(\lambda) - \left(\hat{S}_k(1) - \sum_{i\in\mathcal{K}} h_{ki}\hat{S}_i(1)\right)\right) + \sum_{i\in\mathcal{K}} h_{ki}^2 \mathsf{E}(\hat{S}_i(1))^2 \geq \sum_{i\in\mathcal{K}} h_{ki}^2 \mu_i c_{b,i}^2. \end{aligned} \tag{84}$$

Note, in the second equality, we have applied the following,

$$\mathsf{cov}\left(\left(\hat{\Phi}_k(\lambda) - \left(\hat{S}_k(1) - \sum_{i\in\mathcal{K}} h_{ki}\hat{S}_i(1)\right)\right), \sum_{i\in\mathcal{K}} h_{ki}\hat{S}_i(1)\right) = 0,$$

which is due to the definition of h_{ki} in the above. Observe that the equality in (84) is attained if

$$\hat{\Phi}_k(\lambda t) = \hat{S}_k(t) - \sum_{i\in\mathcal{K}} h_{ki}\hat{S}_i(t). \tag{85}$$

Applying the bound (84) to the stationary queue length in (83), we have

$$\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; S) \geq g(\theta, h) := \sum_{k\in\mathcal{K}} \frac{p_k^2\lambda c_a^2 + \sum_{i\in\mathcal{K}} h_{ki}^2\mu_i c_{b,i}^2}{-2\theta_k}.$$

Hence, the optimal solution to the following optimization problem will give a lower bound for $\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; S)$ over the class $\mathcal{S}$:

$$\min_{\theta_k, h_{ki}} \; g(\theta, h), \quad s.t. \; \sum_{k\in\mathcal{K}} \theta_k = \theta_{\mathcal{K}}, \; \sum_{k\in\mathcal{K}} h_{ki} = 1 \text{ for } i \in \mathcal{K}.$$

Applying the KKT condition, we can find the optimal solution (θ_k^*, h_{ki}^*),

$$\theta_k^* = p_k\theta_{\mathcal{K}}, \; h_{ki}^* = p_k, \quad k, i \in \mathcal{K}. \tag{86}$$

Consequently, a lower bound for $\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; S)$ is given as,

$$g(\theta^*, h^*) = \frac{\lambda c_a^2 + \sum_{k\in\mathcal{K}} \mu_k c_{b,k}^2}{-2\theta_{\mathcal{K}}} \; (= \hat{L}^*). \tag{87}$$

This is indeed the lower bound over the class $\mathcal{D}$ and can be attained if the policy $S \in \mathcal{S}$ induces a limit satisfying

$$\hat{\Phi}_k(\lambda t) = \hat{S}_k(t) - p_k \sum_{i\in\mathcal{K}} \hat{S}_i(t). \tag{88}$$

As motivated by the above discussion, particularly the expression in (85,88), we introduce the so-called service-chasing (SC) policy in the class $\mathcal{S}$ below. Such a routing policy, with properly chosen parameters, will indeed lead to a limit in the form of (88) and achieve the lower bound in (87).

7.1 Service-Chasing Policy and Its Optimality

An SC routing policy for the n-th network is specified with parameters $p^n = (p_k^n)_{k\in\mathcal{K}} \geq 0$ and $h^n = (h_{ki}^n)_{k,i\in\mathcal{K}}$ satisfying

$$\sum_{k\in\mathcal{K}} p_k^n = 1, \text{ and } \sum_{k\in\mathcal{K}} h_{ki}^n = 1 \text{ for } i \in \mathcal{K}, \tag{89}$$

as follows: Upon the ℓ-th arrival ($\ell = 1, 2, \cdots$), pick any server k' such that

$$k' = \arg\min_{k\in\mathcal{K}} \beta_k^n(\ell), \quad \text{with} \tag{90}$$

$$\begin{aligned}\beta_k^n(\ell) &= (\Phi_k^n(\ell-1) - p_k^n\ell) \\ &\quad - \Big[(S_k^n(B_k^n(\Upsilon^n(\ell))) - \mu_k B_k^n(\Upsilon^n(\ell))) - \sum_{i\in\mathcal{K}} h_{ki}^n(S_i^n(B_i^n(\Upsilon^n(\ell))) - \mu_i B_i^n(\Upsilon^n(\ell)))\Big],\end{aligned} \tag{91}$$

and then send the job to that server, i.e.,

$$\Phi_{k'}^n(\ell) = \Phi_{k'}^n(\ell-1) + 1 \;\text{ and }\; \Phi_k^n(\ell) = \Phi_k^n(\ell-1), \; k \neq k'. \tag{92}$$

If we denote $t = \Upsilon^n(\ell)$ and $\tau = \Upsilon^n(\ell-1)$, then, the function $\beta_k^n(\ell)$, defined for choosing server in the above, can be written as:

$$\beta_k^n(\ell) = (\Phi_k^n(E^n(\tau)) - p_k^n E^n(t)) - \Big[(S_k^n(B_k^n(t)) - \mu_k B_k^n(t)) - \sum_{i\in\mathcal{K}} h_{ki}^n (S_i^n(B_i^n(t)) - \mu_i B_i^n(t))\Big]. \quad (93)$$

Note that upon the ℓ-th arrival at time $t = \Upsilon^n(\ell)$, the amount of time that server k has been busy is $B_k^n(t)$, and the server k has served the amount, $S_k^n(B_k^n(t))$, of jobs. Hence, by time t, $(S_k^n(B_k^n(t)) - \mu_k B_k^n(t))$ is the (real-time) deviation of the service completions from its mean (during the busy period before time t), and we call it the "service deviation" for server k. The term inside the squared bracket in (91) then acts as the "target deviation". Thus, by the SC policy the routing deviation will "chase" the target (service) deviation. Observe that when $h_{kk}^n = 1$ and $h_{ki}^n = 0$ $(k \neq i)$ (i.e., the "target" vanishes), the SC policy will also be reduced to an RR policy.

Similar to (74) (for AC policy), we introduce a chasing process under the SC policy, $\Psi^n(t) = (\Psi_k^n(t))_{k\in\mathcal{K}}$:

$$\Psi_k^n(t) := (\Phi_k^n(E^n(t)) - p_k^n E^n(t)) - \Big[(S_k^n(B_k^n(t)) - \mu_k B_k^n(t)) - \sum_{i\in\mathcal{K}} h_{ki}^n (S_i^n(B_i^n(t)) - \mu_i B_i^n(t))\Big]. \quad (94)$$

Note that $\beta_k^n(\ell)$ gives some state information (upon the ℓ-th arrival at time t) excluding the routing information of the arriving job while the process $\Psi^n(t)$ has embodied the routing information of that job. Clearly, at time $t = \Upsilon^n(\ell)$, we have

$$\Psi_k^n(t) = \beta_k^n(\ell) + \phi_k^n(\ell), \quad \sum_{k\in\mathcal{K}} \Psi_k^n(t) = \sum_{k\in\mathcal{K}} \beta_k^n(t) + 1 = 0. \quad (95)$$

In the iterative implementation of the SC policy, we must keep track of the changes in the busy time and the service completion in each server whenever a service is completed or an arrival triggers a routing decision. Consider two consecutive events at times t_1 and t_2, and the chasing process at t_1, $\Psi^n(t_1)$, is known. Denote $\delta_{B,k} := B_k^n(t_2) - B_k^n(t_1)$ and $\delta_{S,k} := S_k^n(B_k^n(t_2)) - S_k^n(B_k^n(t_1))$ for convenience. The busy time changes as $\delta_{B,k} = t_2 - t_1$ if the server k is busy during time interval $[t_1, t_2)$, and $\delta_{B,k} = 0$ otherwise. If there is a service completion at time t_2, say, in server k', then we have $\delta_{S,k'} = 1$, and $\delta_{S,k} = 0$ for $k \neq k'$. The chasing process is updated as:

$$\Psi_k^n(t_2) := \Psi_k^n(t_1) - \Big[\delta_{S,k} - \mu_k \delta_{B,k} - \sum_{i\in\mathcal{K}} h_{ki}^n (\delta_{S,i} - \mu_i \delta_{B,i})\Big].$$

On the other hand, if it is an arrival event at time t_2, say, of the ℓ-th arrival, then $\delta_{S,k} = 0$ for all servers. The arrival process $E^n(t)$ must increase by 1 at time t_2, and we calculate $\beta_k^n(\ell)$ as:

$$\beta_k^n(\ell) := \Psi_k^n(t_1) - p_k^n - \Big[\delta_{S,k} - \mu_k \delta_{B,k} - \sum_{i\in\mathcal{K}} h_{ki}^n (\delta_{S,i} - \mu_i \delta_{B,i})\Big].$$

This determines the routing following (90-92), and thus the chasing process is updated as in (95):

$$\Psi_k^n(t_2) = \beta_k^n(\ell) + \phi_k^n(\ell).$$

Observe that given the parameters (p_k^n, h_{ki}^n, μ_k), the routing decision of an SC policy utilizes the information about the service and routing available up to the decision epoch. And, as we will see in the next theorem, to specify the control parameters (p_k^n, h_{ki}^n) optimally requires a priori knowledge of the model parameters μ_k only. Also note that the routing process and the arrival

process need not be mutually independent due to the busy time process, but they do in the diffusion limit to be established.

Now, apply the SC routing policy defined in (90,91) with the parameters (p^n, h^n) satisfying the requirements in (89). We further require that the condition in (27) is satisfied for some $\theta = (\theta_k)_{k\in\mathcal{K}} < 0$, and

$$h^n_{ki} \to h_{ki}, \quad \text{as } n \to \infty, \text{ for } k, i \in \mathcal{K}, \tag{96}$$

for some $h = (h_{ki})_{k,i\in\mathcal{K}}$ satisfying $\sum_{k\in\mathcal{K}} h^n_{ki} = 1$ for $i \in \mathcal{K}$. Denote such an SC policy sequence as $SC(\{p^n, h^n\}_{n\in\mathcal{N}}, \theta, h)$, or $SC(\theta, h)$ for short. The following theorem shows that the routing policy $SC(\theta, h)$ belongs to $\mathcal{S}$, and can achieve the lower bound performance in (87) over the same class $\mathcal{S}$ as well as the larger class $\mathcal{D}$.

Theorem 6 (a) Suppose the SC policy, $SC(\theta, h)$, is in force, and the initial states satisfy

$$\hat{Q}^n(0) \Rightarrow \hat{Q}(0) \in \mathcal{G}(SC(\theta, h)) := \mathbb{R}^K_+. \tag{97}$$

Then, the weak convergence depicted in (31-35) holds with $\hat{\Phi}_k(t)$ satisfying equation (85) and

$$\hat{X}_k(t) = p_k \hat{E}(t) - \sum_{i\in\mathcal{K}} h_{ki} \hat{S}_i(t) + \theta_k t.$$

Hence, $\hat{Q}_k(t)$ is an RBM with drift θ_k and variation $p_k^2\lambda c_a^2 + \sum_{i\in\mathcal{K}} h_{ki}^2 \mu_k c_{b,k}^2$, and the expected stationary queue lengths are:

$$\mathsf{E}\hat{Q}_k(\infty; SC(\theta, h)) = \frac{p_k^2\lambda c_a^2 + \sum_{i\in\mathcal{K}} h_{ki}^2 \mu_i c_{b,i}^2}{-2\theta_k} \tag{98}$$

$$\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; SC(\theta, h)) = \sum_{k\in\mathcal{K}} \frac{p_k^2\lambda c_a^2 + \sum_{i\in\mathcal{K}} h_{ki}^2 \mu_i c_{b,i}^2}{-2\theta_k}. \tag{99}$$

(b) The routing policy $SC(\theta, h)$ belongs to $\mathcal{S}(\subset \mathcal{D})$, where the approximate routing rates for the n-th system can be given as $p^n = (p^n_k)_{k\in\mathcal{K}}$.
(c) Let (θ, h) be set to (θ^*, h^*) given in (86), or alternatively, choose the sequence of parameters $\{p^n, h^n\}$ as

$$h^n_{ki} = p^n_k = \frac{\mu_k}{\mu_{\mathcal{K}}}, \quad k, i \in \mathcal{K}. \tag{100}$$

Then, the policy $SC^* := SC(\theta^*, h^*)$ is asymptotically optimal in $\mathcal{D}$; that is, for any policy $D \in \mathcal{D}$, we have

$$\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; SC^*) \le \mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; D). \tag{101}$$

The limit satisfies (88) and

$$\hat{X}_k(t) = p_k\Big(\hat{E}(t) - \sum_{i\in\mathcal{K}} \hat{S}_i(t) + \theta_{\mathcal{K}} t\Big) = p_k \hat{X}_{\mathcal{K}}(t). \tag{102}$$

The *postponed state-space collapse* property holds, i.e.,

$$\hat{Q}_k(t) = p_k \hat{Q}_{\mathcal{K}}(t), \quad t \ge \tau, \tag{103}$$

with

$$\tau = \min\Big\{s : -\hat{X}_{\mathcal{K}}(s) = \max_{k\in\mathcal{K}} \frac{\hat{Q}_k(0)}{p_k}\Big\}. \tag{104}$$

The expected stationary queue lengths under SC^* are given in (87), i.e.,

$$\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty; SC^*) = \frac{\lambda c_a^2 + \sum_{k\in\mathcal{K}} \mu_k c_{b,k}^2}{-2\theta_{\mathcal{K}}} \ (= \hat{L}^*), \qquad \mathsf{E}\hat{Q}_k(\infty; SC^*) = p_k \cdot \hat{L}^*. \tag{105}$$

Similar to the proof of Theorem 5, to prove the above theorem, we now employ the chasing property of the SC policy (i.e., the routing deviation chases the service deviation; refer to (91) or (93)) to establish the convergence of the routing processes to the limit given in (85). We again apply the hydrodynamic approach (cf. Lemmas 17 and 18), but with an extra care in handling the busy process embedded in the service process (cf. Lemma 16).

From the above theorem (part (c) in particular), we note that the optimal SC policy does not require estimating the arrival rate. Hence, the policy (just like the JSQ/BR policy) will remain effective when the arrival rate changes over time, which is a desirable feature in real applications.

It would be interesting to observe the SSC properties under various policies here.

Recall from Proposition 2 that the well-known state-space collapse property holds true for all time under the JSQ/BR policy; that is, the convergence $\hat{Q}^n(t) \Rightarrow \hat{Q}(t)$ requires the initial states to (asymptotically) collapse to a one-dimensional space $\mathcal{G}(JSQ)$, and the limit $\hat{Q}(t)$ also then evolves within the space.

In contrast, the state-space collapse property does not exhibit under the RR, PP and AC policies. This is because under these policies, the arrival and routing ($\hat{E}(t)$ and $\hat{\Phi}(t)$) are independent of the service ($\hat{S}(t)$), and therefore the "free process" $\hat{X}(t)$ ($= \hat{\Phi}(t)+p\hat{E}(t)-\hat{S}(t)+\theta t$) evolves in $\mathbb{R}^K_+$. Consequently, the diffusion limit $\hat{Q}(t)$ also evolves in the whole space of $\mathbb{R}^K_+$.

Most interestingly, under the SC policy, the SSC property need not hold for the initial states (see (97)). But the queue length process does collapse to the one-dimensional space $\{q \in \mathbb{R}^K_+ : q_1/\mu_1 = \cdots = q_K/\mu_K\}$ at one time and evolves within the space afterward. In other words, the arrival and routing processes are coupled with the service process, via the mechanism of the SC policy, so that the queue lengths evolve simultaneously and in proportion to the service rates as shown in (103). Such a postponed SSC phenomenon is new in the literature.

8 Estimators and Numerical Studies

Using the diffusion limit theorems established in the above sections, we can derive heavy traffic estimators, or estimators for short, of the performance objective *heuristically*.

From part (a) of Theorem 3, we can turn the result in (52) into the estimators for the expected queue lengths under the RR policy $RR(\theta)$ (i.e., $RR(\{p^n\},\theta)$), by replacing the parameters $(\lambda, c_a, p_k, \theta_k)$ with the pre-limit counterparts $(\lambda^n, c_a^n, p_k^n, n(p_k^n\lambda^n - \mu_k))$ and then "un-scaling" the diffusion scaling through the relationship in (19), for the n-th system:

$$\mathsf{E}Q_k^n(\infty; RR(\theta)) \approx \frac{(p_k^n)^2\lambda^n(c_a^n)^2 + \mu_k c_{b,k}^2}{2(\mu_k - \lambda^n p_k^n)},$$
$$\mathsf{E}Q_{\mathcal{K}}^n(\infty; RR(\theta)) \approx \sum_{k\in\mathcal{K}} \frac{(p_k^n)^2\lambda^n(c_a^n)^2 + \mu_k c_{b,k}^2}{2(\mu_k - \lambda^n p_k^n)}.$$

Recall that the policy $RR(\theta)$ refers to a sequence of RR policies, in which the n-th policy is applied to the n-th system and is specified by the weight parameter p^n satisfying (27). Then, the random variable $Q_k^n(\infty; RR(\theta))$ represents the stationary distribution of $Q_k^n(t)$ under the policy sequence $RR(\theta)$, or more specifically, the n-th policy in the sequence.

For the optimal RR routing policy RR^*, following Theorem 3(c), the weight parameters $p^n = (p_k^n)_{k\in\mathcal{K}}$ can be specified by solving equation (53) described in the theorem. The estimators of the expected stationary queue lengths, for the server k and the system, are

$$\mathsf{E}Q_{\mathcal{K}}^n(\infty; RR^*) \approx \frac{\left(\sum_j \sqrt{(p_j^n)^2\lambda^n(c_a^n)^2 + \mu_j c_{b,j}^2}\right)^2}{2(\mu_{\mathcal{K}} - \lambda^n)} := L(RR^*), \tag{106}$$
$$\mathsf{E}Q_k^n(\infty; RR^*) \approx \frac{\sqrt{(p_k^n)^2\lambda^n(c_a^n)^2 + \mu_k c_{b,k}^2}\left(\sum_j \sqrt{(p_j^n)^2\lambda^n(c_a^n)^2 + \mu_j c_{b,j}^2}\right)}{2(\mu_{\mathcal{K}} - \lambda^n)}.$$

As indicated by the weak convergence in Theorem 3(a), the closer the arrival rate λ^n is to the total service rate $\mu_{\mathcal{K}}$ (i.e., the larger the index n is), the more accurate the above estimators must be.

Similarly, from Theorem 5, part (a) in particular, an estimator under the AC policy $AC(\theta, h)$ (i.e., $AC(\{p^n, h^n\}_{n\in\mathcal{N}}, \theta, h)$), with the parameters (p^n, h^n) satisfying in (70), can be written as

$$\mathsf{E}Q_k^n(\infty; AC(\theta,h)) \approx \frac{(h_k^n + p_k^n)^2\lambda^n(c_a^n)^2 + \mu_k c_{b,k}^2}{2(\mu_k - \lambda^n p_k^n)},$$
$$\mathsf{E}Q_{\mathcal{K}}^n(\infty; AC(\theta,h)) \approx \sum_{k\in\mathcal{K}} \frac{(h_k^n + p_k^n)^2\lambda^n(c_a^n)^2 + \mu_k c_{b,k}^2}{2(\mu_k - \lambda^n p_k^n)}.$$

And from part (c), an estimator under the optimal AC policy AC^* with the parameters (p^n, h^n) specified in (80), can be written as

$$\mathsf{E}Q_{\mathcal{K}}^n(\infty; AC^*) \approx \frac{\lambda^n(c_a^n)^2 + \left(\sum_j \sqrt{\mu_j c_{b,j}^2}\right)^2}{2(\mu_{\mathcal{K}} - \lambda^n)} := L(AC^*), \tag{107}$$
$$\mathsf{E}Q_k^n(\infty; AC^*) \approx (h_k^n + p_k^n)L(AC^*) = \frac{\sqrt{\mu_k c_{b,k}^2}}{\sum_{j\in\mathcal{K}}\sqrt{\mu_j c_{b,j}^2}} L(AC^*).$$

From part (a) of Theorem 6, an estimator under the SC policy $SC(\theta, h)$ (i.e., $SC(\{p^n, h^n\}_{n\in\mathcal{N}}, \theta, h)$), with the parameters (p^n, h^n) satisfying (89), can be written as:

$$\mathsf{E}Q_k^n(\infty; SC(\theta,h)) \approx \frac{(p_k^n)^2\lambda^n(c_a^n)^2 + \sum_{i\in\mathcal{K}}(h_{ki}^n)^2\mu_i c_{b,i}^2}{2(\mu_k - \lambda^n p_k^n)},$$
$$\mathsf{E}Q_{\mathcal{K}}^n(\infty; SC(\theta,h)) \approx \sum_{k\in\mathcal{K}} \frac{(p_k^n)^2\lambda^n(c_a^n)^2 + \sum_{i\in\mathcal{K}}(h_{ki}^n)^2\mu_i c_{b,i}^2}{2(\mu_k - \lambda^n p_k^n)}.$$

And from part (c), an estimator under the optimal SC policy SC^* with the parameters (p^n, h^n) specified in (100), can be written as:

$$\mathsf{E}Q_{\mathcal{K}}^n(\infty; SC^*) \approx \frac{\lambda^n(c_a^n)^2 + \sum_k \mu_k c_{b,k}^2}{2(\mu_{\mathcal{K}} - \lambda^n)} := L^*, \quad \mathsf{E}Q_k^n(\infty; SC^*) \approx \frac{\mu_k}{\mu_{\mathcal{K}}} L^*. \tag{108}$$

For comparison, we also write the estimator under the JSQ (or BR) policy using Proposition 2 (cf. (47,48)) as follows,

$$\mathsf{E}Q_{\mathcal{K}}^n(\infty; JSQ) \approx \frac{\lambda^n(c_a^n)^2 + \sum_{k\in\mathcal{K}} \mu_k c_{b,k}^2}{2(\mu_k - \lambda^n p_k^n)} = L^*, \quad \mathsf{E}Q_k^n(\infty; JSQ) \approx \frac{1}{K} L^*. \tag{109}$$

Clearly, the estimator of the expected stationary queue length under the SC policy attains the lower bound estimate as the JSQ policy does. (We have not discussed the PP policy in this section since it has been widely studied in the literature.)

In our study, we have assumed prior (static) information about key parameters (i.e., first two moments of the arrival and service processes — λ^n, c_a^n, μ_k, and $c_{b,k}$), which may require effort to determine in many applications. It is interesting to observe that an optimal routing policy requires specifying some of these parameters a priori but is insensitive to the others. Specifically, when no (dynamic) state information is available, determining the optimal RR policy requires all the first two moments (cf. (3,53)). In contrast, the JSQ policy uses none of these parameters; the queue length state synthesizes the arrival and service state information for purpose of routing control. The optimal AC and SC policies lie in between in terms of

the "usage" of these parameters. That is, using the state information of the arrival ($E^n(t)$), the optimal AC policy depends on λ^n, μ_k and $c_{b,k}$, but not c_a^n (cf. (73,80)); while with the state information of the service ($S_k^n(t)$), the optimal SC policy requires a less amount of prior information, i.e., the mean of service times (μ_k) only (cf. (93,100)).

In application, the above estimators apply to a specific, original heavily-loaded system, without artificially introducing a system sequence and the resulting diffusion limit. Thus, the parameters like λ^n, ρ^n, p^n and h^n refer to those in the original system of interest, and the index n is not required. In the reminder of this section, we omit the index n to lighten the burden of notation.

To carry out the simulation studies, we consider a system with five servers. The service times in each server follow an exponential distribution with service rates $(\mu_1, \cdots, \mu_5) = (0.1, 0.1, 0.2, 0.2, 0.4)$ and total rate $\mu_{\mathcal{K}} = 1$. Jobs arrive following a Poisson process with arrival rate λ. The coefficients of variation are then $c_a^2 = c_{b,1}^2 = \cdots = c_{b,5}^2 = 1$. The system traffic intensity is $\rho = \lambda$ $(= \lambda/\mu_{\mathcal{K}})$, and this parameter will vary in the numerical studies.

First, we examine the accuracy of the estimators by comparing it with the simulated performances under various policies. For each of the routing policies (i.e., RR^*, AC^*, SC^* and JSQ), we compute the expected stationary queue length for the system traffic intensity $\rho = 0.80, 0.81, \cdots, 0.99$, by using the respective estimators in (106,107,108,109) (the "estimated" performance) and by simulation (the "simulated" performance). By adjusting the length and the number of simulation runs, we ensure the width of 99% confidence interval is always within 1% of the mean for all the simulation results reported in the paper. Thus, we use the simulated mean total queue length performance as the proxy of the (theoretical) expected performance.

In the comparisons shown in Figures 2-5, the estimated mean total queue length (continuous/red curves) generally approximate the simulated performance (dashed/blue curves) more closely when the traffic intensity ρ approaches 1, which is predicted by the associated heavy traffic results in Theorems 3, 5 and 6 and Proposition 2, respectively. For the optimal RR policy RR^*, as indicated by the dot/green line in Figure 2, the estimator yields an over estimation of the expected stationary queue length as compared with the simulated performance. The ratio of the estimated to the simulated performance, i.e., (the proxy of) $L(RR^*)/\mathsf{E}Q_{\mathcal{K}}(\infty; RR^*)$, ranges from 1.17 to 1.09 for "moderate" traffic intensity $\rho = 0.80, \cdots, 0.90$, and then drops to 1.036, 1.030 and 1.024 for $\rho = 0.97$, 0.98 and 0.99. For the optimal AC policy AC^*, we observe in Figure 3 that the estimator first over-estimates and then gets very close to the expected stationary queue length. The performance ratio (dot/green line), $L(AC^*)/\mathsf{E}Q_{\mathcal{K}}(\infty; AC^*)$, decreases from 1.14 to 1.00 when ρ increases from 0.80 to 0.99.

Both estimators for the optimal SC and the JSQ policies under-estimate the expected performances, as indicated by the performance ratios (dot/green line) in Figures 4 and 5, respectively. The performance ratios, $L^*/\mathsf{E}Q_{\mathcal{K}}(\infty; D)$ $(D = SC^*, JSQ)$, raise from 0.53 to 0.98 for the optimal SC policy and from 0.57 to 0.96 for the JSQ policy, when ρ increases from 0.80 to 0.99. We should note that, though both policies achieve the optimal performance assymptotically, they require different parameters and state information in implementation. Moreover, when the realtime queue length information is available, the JSQ policy is often easy to implement and does not require a priori knowledge of system parameters.

To close this section, we deviate from the heavy traffic analysis slightly, and present some preliminary simulation studies for systems under varying traffic loads. We aim to shed light on the selection of routing policy when the traffic intensity, variability of interarrival and service times, and number of servers vary in application, keeping in mind that different policies may employ different state information and apply in different settings. We will look at the average waiting time (including both queueing and service), instead of the average queue length, for a different performance perspective.

It is known that the JSQ policy may not be optimal in the usual sense (rather than asymptotically) when the servers are not identical ([57]). Thus, when real-time queue length information

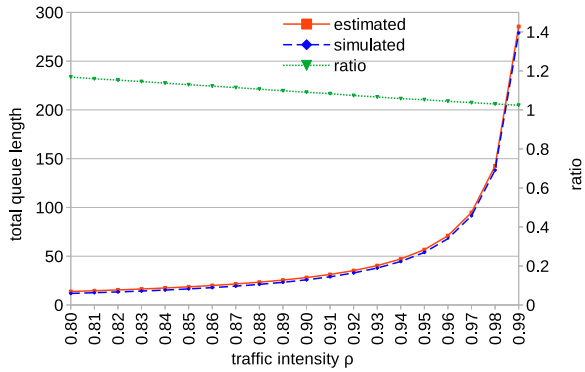

Figure 2: RR policy

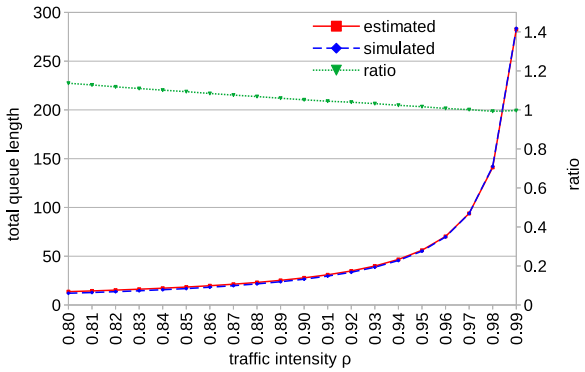

Figure 3: AC policy

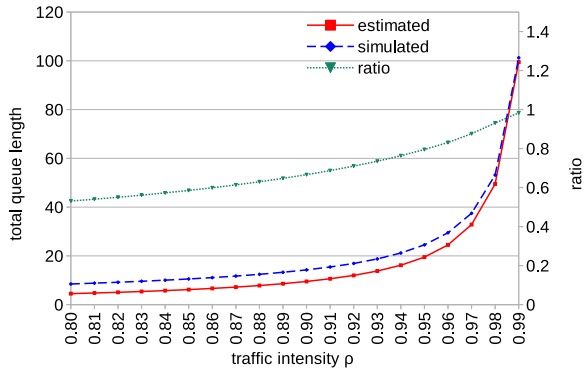

Figure 4: SA policy

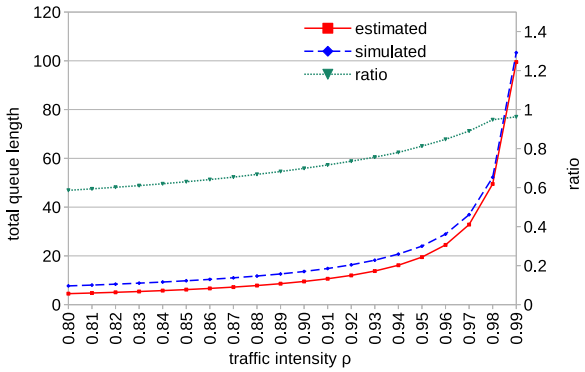

Figure 5: JSQ policy

is available, many policies have been proposed to improve the performance (e.g., [4,49]). Hence, we include some of those policies in the simulations and describe them following Banawan and Zeidat [4].

- The shortest expected delay policy (SED) tries to minimize the expected waiting time of a job by dispatching the job, upon its arrival at time t, to the server $k' = \arg\min_k \frac{q_k+1}{\mu_k}$. Here, $q_k := Q_k(t-)$ is the number of jobs at server k (including the one being served, if any) immediately prior to the time t.
- The never queue policy (SEDNQ) is a variation of the SED policy. It assigns a new arrival to an idle server, if any; and if there are more than one idle server, assign the job to the fastest server. On the other hand, if all servers are busy, the policy behaves similar to the SED policy.
- The greedy throughput policy (GTP) aims at maximizing the (approximate) system throughput by assigning an arrival to the server $k' = \arg\max_k (\frac{\mu_k}{\mu_k+\lambda})^{q_k+1}$.
- The adaptive separable policy (ASP) provides another improvement over the SED policy by adjusting the service rate of the server based on its traffic intensity (utilization) in estimating the expected delay. It dispatches the incoming job to the server $k' = \arg\min_k \frac{q_k+1}{\mu_k(1-\rho_k)}$. Here, the traffic intensity of server k, ρ_k, depends on the routing policy and is estimated and updated by keeping track of the server's busy time in implementation.

We continue to examine the system with five servers considered just above, and compare the average waiting times of different policies as the traffic intensity changes from $\rho = 0.10$ to 0.98 for four different cases in Figures 6-9. We assume a Poisson arrival process in all cases. In the first case (Figure 6), the service times are exponentially distributed with the same rates $(\mu_1, \cdots, \mu_5) = (0.1, 0.1, 0.2, 0.2, 0.4)$ as above. In the second and third case (Figures 7 and 8), we consider a more realistic service time distribution, the Gamma distribution, with the

same service rates. The service times in the second case have a large coefficient of variation for all servers (with $c_{b,k}^2 = 10$). In contrast, in the third case, the servers have imbalanced variabilities. Specifically, a slow server has a coefficient of variation much larger than the others (with $c_{b,1}^2 = 10$, and $c_{b,k}^2 = 0.1$ for $k = 2, 3, 4, 5$). In the fourth case (Figures 9), we replicate the five-server model by 10 times and merge the arrival streams of the 10 identical models to form a single Poisson arrival stream. Same as the first case, all service times are exponentially distributed. Thus, we examine a many-server system with 50 servers in total.

From the simulations, we can see that the RR^* policy, without using any state information, may be comparable to (or even better than) certain policies that use state information for systems under light or moderate traffic intensity. While the AC^* and SC^* policies are asymptotically optimal (Theorems 5 and 6), their behaviors look more complicated when ρ drops to moderate or light traffic. For example, the AC^* policy outperforms the RR^* policy apparently in the third case (Figure 8) when the system is heavily loaded. However, its advantage diminishes when the traffic intensity ρ drops below 0.9, and its performance even falls behind the RR^* policy when ρ moves from 0.88 to 0.80. Figure 9 shows that when there are many servers, the performance of the SC^* policy degrades significantly even when the traffic intensity is close to one. This highlights the need for many-server regime in the study of large-scale systems again, as pointed out in the Introduction.

When the real-time queue length information is available, none of the policies dominate all the others for the four cases. Nevertheless, it appears that the GTP policy performs well when the traffic intensity, server heterogeneity and number of servers vary. In contrast, the JSQ policy does not require estimating any model parameter, but yields a waiting time significantly longer than the GTP policy. The JSQ policy is inferior to the SEDNQ policy too; this is because, while both policies enforce the use of idle servers, the SEDNQ policy incorporates the service rates to speed up services. The SED policy appears less reliable and performs worse when the traffic intensity and the number of servers increase, as illustrated in Figure 9. As a variation of SED, the SEDNQ policy yields a more reliable performance though it is less favorable for highly heterogeneous servers under light traffic (cf. [4]). Indeed, further investigation is required to understand when the "no queue" strategy gives improvement and whether the strategy can be applied to improve the other routing policies (cf. [23]). Another variation of SED, the ASP policy often gives a good performance comparable to the GTP policy, but similar to SED in Figure 9, appears less reliable when the traffic intensity and the number of servers increase.

In summary, when moving beyond the heavy traffic setting considered in this paper, the performance of a routing policy depends on multiple factors in a complex manner, e.g., the traffic intensity ρ, the number of servers, the heterogeneity of servers (in service rates, variability and distribution of service times), tie-breaking method of the policy, and even the order of servers. Therefore, further research about the efficient use of various state information in the non-heavy traffic setting is required.

9 Interchange of Limits

We have established process-wise convergences in diffusion limit theorems, Theorems 3, 4, 5 and 6, under various routing policies, which indicate that the limit $\hat{Q}(t)$ is "close" to the pre-limit process $\hat{Q}^n(t)$ for all finite time ($t < \infty$). Presuming that their stationary performances (at time $t = \infty$) are also close to each other, we use $\mathsf{E}\hat{Q}_{\mathcal{K}}(\infty)$ as an approximation of $\mathsf{E}\hat{Q}^n_{\mathcal{K}}(\infty)$ to derive the heavy traffic estimators heuristically in the previous section.

The foregoing presumption can be justified more rigorously by studying the *interchange of limits*. This idea is illustrated more formally by the rectangle in Figure 10, a setup due originally to Gamarnik and Zeevi [24]. We take the RR policy as an example and follow Ye and Yao [67] to describe the idea.

First, in Theorem 3(a), we have established the diffusion limit under the heavy traffic condi-

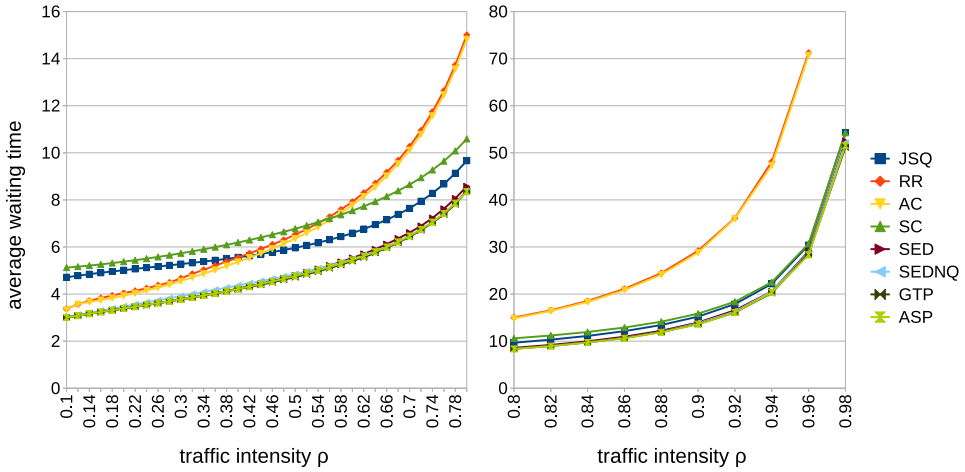

Figure 6: Exponential service times

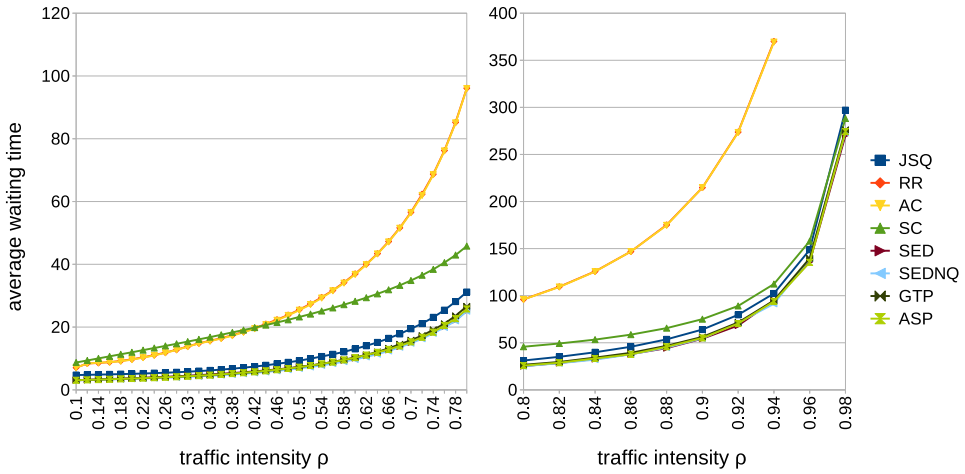

Figure 7: Gamma service, high variability

tion, "$\hat{Q}(t) = \lim_{n\to\infty} \hat{Q}^n(t)$", which is the task designated to the left vertical side, edge I, of the rectangle. (Equations in quotation marks give an intuitive description rather than a rigorous formulation.) Next, for each n, we want to claim that $\hat{Q}^n(t)$ has a stationary distribution as $t \to \infty$, with $\hat{Q}^n(\infty)$ denoting the random variable associated with this limiting distribution. This step is represented on edge II of the rectangle, and can be accomplished by applying the fluid model approach developed by Rybko and Stolyar [48], Dai [17], Chen [13] and Stolyar [51], etc. It will be a by-product of establishing edge IV in this section. Analogously, as represented by the edge III, Theorem 3(a) also implies that the diffusion limit $\hat{Q}(t)$ has a stationary distribution, embodied by $\hat{Q}(\infty)$. The diffusion approximation is then to use this last stationary distribution, that of the diffusion limit $\hat{Q}(t)$, as an approximation for the stationary distribution of the queue length in the original network. This is tantamount to claiming weak convergence on edge IV,

$$\text{“}\lim_{n\to\infty} \hat{Q}^n(\infty) = \hat{Q}(\infty)\text{”}, \quad \text{or} \quad \text{“}\lim_{t\to\infty}\lim_{n\to\infty} \hat{Q}^n(t) = \lim_{n\to\infty}\lim_{t\to\infty} \hat{Q}^n(t)\text{”}. \tag{110}$$

Thus, to justify the diffusion approximation boils down to justifying the interchange of two limits, $n \to \infty$ and $t \to \infty$.

From the above description, it remains to establish edges II and IV of the rectangle, particularly the convergence of the stationary distributions, and furthermore the stationary moments,

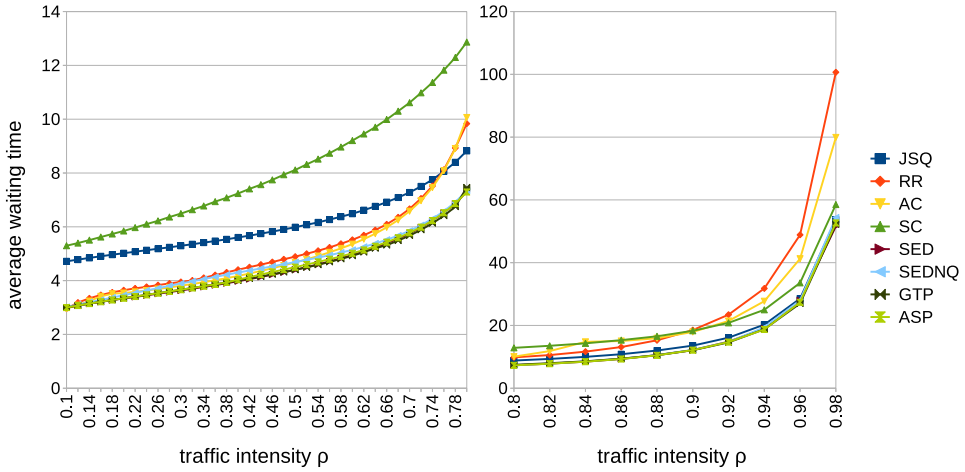

Figure 8: Gamma service, imbalance variability

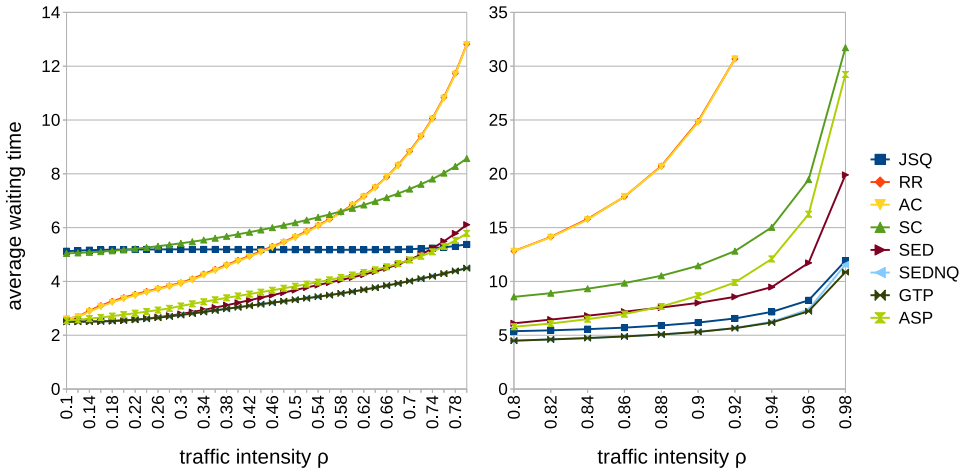

Figure 9: Exponential service, 50 servers

of pre-limit queue lengths to those of the diffusion limit for the routing policies being studied. Recall that the RR policy can be viewed as a special case of the AC or SC policy. In addition, our parallel system under the PP routing is a special case of the generalized Jackson network, for which the interchange-of-limits problem has been thoroughly studied in [11]. Hence, the following discussions focus on the AC, SC, JSQ and BR policies, with the RR policy treated as a special case.

To establish the interchange of limits, we will modify the sequence of systems to allow more general initial states, construct a Markovian representation for the systems, and introduce additional conditions on distributions of interarrival and service times.

First, we allow the initial residual arrival and service times, u_1^n and $v_{k,1}^n$, to be any nonnegative numbers while the rest of the interarrival time and service time sequences, $\{u_\ell^n, \ell \geq 2\}$ and $\{v_{k,\ell}^n, \ell \geq 2\}$, remain mutually independent i.i.d. random sequences. The arrival and service processes, $E^n(t)$ and $S_k^n(t)$, are then *delayed* renewal processes. Moreover, we include the "initial surplus" of routing deviation to the chasing processes in (74) and (94), for AC policy,

$$\Psi_k^n(t; AC(\theta, h)) := \Psi_k^n(0) + (\Phi_k^n(E^n(t)) - p_k^n E^n(t))) - h_k^n(E^n(t) - \lambda^n t), \qquad (111)$$

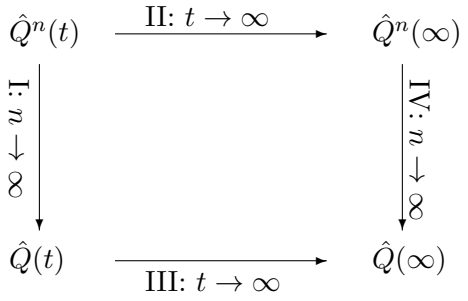

Figure 10: Interchange of limits

and for SC policy,

$$\Psi_k^n(t; SC(\theta, h)) := \Psi_k^n(0) + (\Phi_k^n(E^n(t)) - p_k^n E^n(t))) - \left[(S_k^n(B_k^n(t)) - \mu_k B_k^n(t)) - \sum_i h_{ki}^n (S_i^n(B_i^n(t)) - \mu_i B_i^n(t)) \right]. \tag{112}$$

Under either AC or SC policy, the following initial condition is imposed,

$$\sum_{k \in \mathcal{K}} \Psi_k^n(0) = 0. \tag{113}$$

For JSQ or BR policy, the chasing process is not necessary, and therefore for ease of presentation, we denote

$$\Psi_k^n(t; D) \equiv 0, \quad \text{for } D = JSQ, BR. \tag{114}$$

We follow the standard approach (e.g., [11, 17, 24, 25]) to construct a Markov process representation of the network by appending supplement information to the queue length state. First, we denote the residual interarrival and service times (at each time instant) as $U^n(t)$ and $V^n(t) = (V_k^n(t))_{k \in \mathcal{K}}$, $t \geq 0$, where:

$$U^n(t) = \sum_{\ell=1}^{E^n(t)+1} u_\ell^n - t, \qquad V_k^n(t) = \left[\sum_{\ell=1}^{S_k^n(B_k^n(t))+1} v_{k,\ell}^n - B_k^n(t) \right] \cdot 1_{\{Q_k^n(t)>0\}}. \tag{115}$$

That is, at any given time t, $U^n(t)$ is the remaining time before the next arrival, and $V_k^n(t)$ is the remaining service time for the class-k job that is in service. (If there is no class-k job and the server k is idle, $V_k^n(t)$ is the "remaining" service time for the class-k job that has just left, i.e., $V_k^n(t) = 0$.) Observe that at time $t = 0$, we have $U^n(0) = u_1^n$ and $V_k^n(0) = v_{k,1}^n$ (if the server k is busy), the residuals at time zero introduced above. (Note, the residual service time process defined here is a slight refinement of those in previous studies (e.g., [11, 17]) by introducing the non-idling component, $1_{\{Q_k^n(t)>0\}}$.) Hence, below we shall refer to $U^n(t)$ and $V_k^n(t)$ as "residuals" (at time t) as well.

Then, $\Xi^n(t) = (Q^n(t), U^n(t), V^n(t), \Psi^n(t))$ is a strong Markov process, taking values on the nonnegative orthant of the $(3K + 1)$-dimensional real space, denoted by $\mathcal{X}$ (cf. [17, 21, 36]). Clearly, the dynamics of the Markov process $\Xi^n(t)$ will be completely determined when the initial state $\Xi^n(0)$ is given. Below, we will often consider many copies of the same network, each starting from a different initial state. To highlight the dependence on the initial state, we will append it to the argument of the corresponding Markov process and queue length process. Hence, instead of $\Xi^n(t)$ and $Q^n(t)$, we will write $\Xi^n(t; x)$ and $Q^n(t; x)$, with $\Xi^n(0) = x \in \mathcal{X}$ being the initial state.

The above Markov representation of the network is necessary for much of the proofs below, which rely heavily on the theory of Markov processes. It would be useful, however, to keep in mind that in the special case of Poisson arrivals and exponential service times, the queue length $Q^n(t)$, plus the chasing process $\Psi^n(t)$ if the routing policy is AC or SC, already constitute a Markov process, instead of the more elaborate $\Xi^n(t)$. Focusing on this special case, as the reader may choose to do below, has the advantage of getting directly to the main ideas, without interference from all the technicalities involving the appended states $U^n(t)$ and $V^n(t)$.

As previously stated, we require the primitives of the network, the interarrival and service times, to possess a finite second moment. In the following, we justify not only the convergence of stationary distributions but also the convergence of stationary moments. This requires strengthening the second moment to a higher (p-th) moment condition that holds *uniformly* for all the systems. To avoid technicality, we assume that the system sequence is driven by the same primitives except the initial arrival and service times; that is, assume for all $n \in \mathcal{N}$,

$$\lambda^n u_\ell^n = \lambda^1 u_\ell^1 \ \text{ and } \ v_{k,\ell}^n = v_{k,\ell}^1, \quad \ell \geq 2,\ k \in \mathcal{K}. \tag{116}$$

The *p-th moment condition* reads: for a given $p > 2$, assume all interarrival and service times have uniform bounded p-th moments, i.e.,

$$\sup_{n\in\mathcal{N}} \mathsf{E}\left[(u_\ell^n)^p + \sum_{k\in\mathcal{K}} (v_{k,\ell}^n)^p\right] < \infty, \quad \text{or} \quad \mathsf{E}\left[(u_2^1)^p + \sum_{k\in\mathcal{K}} (v_{k,2}^1)^p\right] < \infty. \tag{117}$$

Clearly, this strengthens Bramson's uniform second moment condition in (12). As we will see, we need $p > m + 1$ when the convergence of the m-th moment of the queue length is required. And, to justify our (first moment) heavy-traffic estimators as valid stationary approximations amounts to requiring $m = 1$ and thus a uniform bounded $(2 + \epsilon)$-th moment condition.

In addition, we assume that for all n and $\ell \geq 2$,

$$\mathsf{P}\{u_\ell^n \geq a\} > 0, \quad \text{for any } a > 0; \tag{118}$$

and that for some integer $j \geq 2$ and some nonnegative function $p(x)$ satisfying $\int_0^\infty p(s) > 0$, the following inequality holds:

$$\mathsf{P}\left\{a \leq \sum_{\ell=2}^{j} u_\ell^n \leq b\right\} \geq \int_a^b p(x)dx, \quad \text{for any } 0 \leq a < b. \tag{119}$$

These are certain forms of "spread-out" condition, required to guarantee the positive (Harris) recurrence and hence the uniqueness of the stationary distribution of the pre-limit systems in edge II of Figure 10. They also appeared in prior works, e.g., [8, 17].

Moreover, to avoid non-essential technicalities, we assume that the interarrival times $\{u_\ell^n\}$ are continuous random variables. Consequently, the time of a job's arrival will not coincide with the time of any service completion or any other arrival almost surely. (It should be noted that to remove this assumption involves a tedious discussion about the possible simultaneous events of arrivals and service completions. This can be done since at any time of job arrival, the number of simultaneous arrivals follows a geometric distribution if we allow $u_\ell^n = 0$ with a positive probability, and this number has a moment of any order. We leave a formal discussion to the interested reader.)

Given the Markovian state descriptor and additional conditions on the interarrival and service times described above, we are ready to study the interchange-of-limits problem.

For edge II in Figure 10, we have the following theorem.

Theorem 7 Consider the sequence of systems $\{\hat{\Xi}^n(t)\}_{n\in\mathcal{N}}$ under the $RR(\theta)$, $AC(\theta, h)$, $SC(\theta, h)$, JSQ or BR policy. For any sufficiently large n, $\hat{\Xi}^n(t) = (\hat{Q}^n(t), \hat{U}^n(t), \hat{V}^n(t), \hat{\Psi}^n(t))$ is positive

recurrent and has a unique stationary distribution. Furthermore, if the p-th moment condition in (117) also holds, the stationary queue length has a finite $(p-1)$-th moment and

$$\lim_{t\to\infty} \mathsf{E}|\hat{Q}^n(t;x)|^{p-1} = \mathsf{E}|\hat{Q}^n(\infty)|^{p-1} < \infty, \quad \text{for any initial state } \hat{\Xi}^n(0) = x, \tag{120}$$

where $\hat{Q}^n(\infty)$ stands for a random variable (vector) following the stationary distribution of $\hat{Q}^n(t)$.

Note that this theorem holds under the heavy-traffic condition in (11) with $\theta_{\mathcal{K}} < 0$ in particular, which is enforced throughout the paper and guarantees the usual traffic condition, $\lambda^n < \mu_{\mathcal{K}}$, for sufficiently large n. The second moment condition in (12) is used to ensure the positive recurrence of $\hat{\Xi}^n(t)$ while the "furthermore" part of the theorem requires the stronger p-th moment condition.

It should be pointed out that the stability of our parallel server system under the JSQ/BR policy, a conclusion in the above theorem, is known in the literature (cf. [10, 16, 38]). For example, Bramson [10] studied stability of the JSQ policy for systems with sophisticated service disciplines, which implies the stability of JSQ for the parallel server model being studied. In Chen and Ye [16], the fluid model associated with the JSQ/BR policy (without initial residuals) is well-understood, and thus the stability and the positive recurrence of the system with JSQ/BR, as stipulated in Theorem 7, can be inferred immediately.

We establish the above theorem following the fluid model approach that relates the stability of a queueing (network) system to the stability of the associated fluid model. To apply the approach, we first identify the fluid model corresponding to each of the pre-limit systems (Lemma 19(a)). Intuitively, it is obtained by replacing the random arrival and service processes by deterministic fluid flows. Then, for each (the n-th) system $\hat{\Xi}^n(t)$, we show that the corresponding fluid model, mainly the queue length analogy $\hat{q}^n(t)$, is stable (Lemma 19(b)). That is, starting with any initial state bounded by 1, the queue length $\hat{q}^n(t)$ drains to zero by a time t_0 that is independent of k. Once these steps are accomplished, the results of [17, 19] can be invoked to complete the proof of the above theorem.

Notably, the above stability property is uniform in the sense that the time t_0 can be specified independent of the (sufficiently large) index n. Such a *uniform stability* will be used as an important tool in establishing the edge IV below. For more general stochastic processing networks, we have shown that such a uniform stability property is guaranteed by the stability of fluid model corresponding to the diffusion limit (cf. [66]). Those networks may involve more complicated features such as the complex network topology, feedback mechanism, and current resources. That approach, somewhat detoured, can be applied to establish the uniform stability for our model. Nevertheless, as our (pre-limit) parallel server systems involve simple fluid models only, which are one-dimensional reflecting mappings, we are able to establish the property directly.

Now, we are ready to establish our main results for the edge IV, the convergence of stationary distributions and moments, in the following two theorems.

Theorem 8 Consider the sequence of systems $\{\hat{\Xi}^n(t)\}_{n\in\mathcal{N}}$ under the $RR(\theta)$, $AC(\theta,h)$, $SC(\theta,h)$ or JSQ policy. Then, the following weak convergence holds:

$$\hat{Q}^n(\infty) \Rightarrow \hat{Q}(\infty), \quad \text{as } n\to\infty.$$

Furthermore, if the p-th moment condition in (117) also holds, we have, for any $m \in [0, p-1)$,

$$\lim_{n\to\infty} \mathsf{E}|\hat{Q}^n(\infty)|^m = \mathsf{E}|\hat{Q}(\infty)|^m.$$

Theorem 9 Consider the sequence of systems $\{\hat{\Xi}^n(t)\}_{n\in\mathcal{N}}$ under the BR (with $c \geq 2$) policy. Then, the following weak convergence holds:

$$\hat{Q}^n(\infty) \Rightarrow \hat{Q}(\infty), \quad \text{as } n \to \infty.$$

Furthermore, suppose for some $p^* > 2(p+2)$, a *p^*-th moment condition* is in force: all interarrival and service times have bounded p^*-th moments,

$$\sup_{n\in\mathcal{N}} \mathsf{E}\left[(u_\ell^n)^{p^*} + \sum_{k\in\mathcal{K}} (v_{k,\ell}^n)^{p^*}\right] < \infty. \tag{121}$$

Then, for any $m \in [0, p-1)$, we have:

$$\lim_{n\to\infty} \mathsf{E}|\hat{Q}^n(\infty)|^m = \mathsf{E}|\hat{Q}(\infty)|^m.$$

In [66, 67], we developed a recipe for studying the interchange-of-limits problem for a wide range of stochastic processing networks under heavy traffic. Here, we adopt the recipe, by carefully handling the additional features due to the routing control mechanisms, to prove the above two theorems.

As we mention in the introduction, the crucial step in the recipe is to bound the p-moment of the state process, which is mainly to bound $\mathsf{E}\sup_{0\leq s\leq t}|\hat{Q}^n(s)|^p$ (cf. (209,266)). For the AC, SC and JSQ policies, we will do so by verifying a pathwise bound, meaning that the queue length process $\hat{Q}^n(t)$ can be bounded by the free process comprising primitive processes $\hat{E}^n(t)$ and $\hat{S}^n(t)$ along with the initial state (Lemma 23). This is possible because under the AC and SC policies, the routing process $\hat{\Phi}^n(t)$ "chases" the (scaled) arrival and the service processes, and thus can be bounded by the respective processes. For the JSQ policy, the pathwise bound is established through a delicate analysis of the sample path. As the processes $\hat{E}^n(t)$ and $\hat{S}^n(t)$ has a moment bound, the pathwise bound then yields the required p-moment bound (Lemma 24).

However, for the BR policy, it is not obvious how to establish a similar pathwise bound. To overcome this difficulty, we assume a higher moment condition, the p^*-th moment of the primitives (cf. (121)). Under this condition, we are able to identify and focus on a sequence of "regular" events, in which the state processes behave "nicely," and the probabilities of these events occurring approach 1 at a certain rate (cf. Lemma 28). We then apply the hydrodynamic approach to derive a pathwise bound for the queue length process for sample paths in regular events (cf. Lemmas 29 and 30). Therefore, the queue length processes, when restricted to the regular events, possess a bounded p-th moments. On the other hand, the p-th moment of the queue length processes, restricted to the non-regular events, have the same bound owing to the small probability of such events. Combining these two cases leads to the desired p-th moment of queue length processes (cf. (266)).

Once the p-th moment bound is established for queue length processes, along with the uniform stability established on edge II, we can prove the uniform p-th moment stability of the queue length processes (cf. (212,267)), which will lead to the tightness of $\{\hat{Q}^n(\infty), n \in \mathcal{N}\}$ and the convergence of stationary distributions and moments on edge IV, following the approach (by now standard) in [11, 19, 66, 67].

A Appendix

A.1 Preliminary: One-Dimensional Reflection Mapping, Reflecting Brownian Motion, Law Of Large Numbers

We first collect some useful results about the reflection mapping and reflecting Brownian motion. (Note, notation introduced in the following three lemmas, except the mappings Φ and Ψ, are valid only within this subsection, and should not be confused with those elsewhere in the paper.)

Lemma 10 (e.g., [15], Theorem 6.1) Let $x(t)$, $t \geq 0$, be any one-dimensional real-valued function that is right-continuous and with left limits (RCLL). Then, there exists a unique pair of RCLL functions $(y(t), z(t))$ satisfying

$$z(t) = x(t) + y(t) \geq 0, \tag{122}$$

$$y(t) \text{ is non-decreasing in } t, \text{ with } y(0) = 0, \tag{123}$$

$$\int_0^\infty z(t)dy(t) = 0. \tag{124}$$

In fact, the unique pair $(y(t), z(t))$ can be expressed as

$$y(t) = \sup_{0 \leq s \leq t} [-x(s)]^+, \tag{125}$$

$$z(t) = x(t) + \sup_{0 \leq s \leq t} [-x(s)]^+. \tag{126}$$

We call the functions $y(t)$ and $z(t)$ the regulator and reflected process of the function ("free" process) $x(t)$, respectively. Hence, the relationships in (122-124) define a one-dimensional reflection mapping, known as the *Skorohod mapping*, denoted as: $(z, y) = (\Phi(x), \Psi(x))$. The mappings Φ and Ψ are Lipschitz continuous under the uniform topology (e.g., [15], Exercises 1 and 2 of Chapter 6): for any RCLL functions x and x', the following inequalities hold for any $t \geq 0$,

$$\sup_{0 \leq s \leq t} |\Psi(x)(s) - \Psi(x')(s)| \leq \sup_{0 \leq s \leq t} |x(s) - x'(s)|, \tag{127}$$

$$\sup_{0 \leq s \leq t} |\Phi(x)(s) - \Phi(x')(s)| \leq 2 \sup_{0 \leq s \leq t} |x(s) - x'(s)|. \tag{128}$$

Let $X(t)$, $t \geq 0$, be a one-dimensional Brownian motion that starts at $X(0)$, and has drift θ and standard deviation σ. Then, we call the process $Z(t) := \Phi(X)(t)$, $t \geq 0$, a one-dimensional *reflecting Brownian motion* (RBM), which we characterize in the lemma below.

Lemma 11 (e.g., [15], Theorem 6.2) The RBM $Z(t)$ has a stationary distribution if and only if $\theta < 0$, in which case the stationary distribution is exponential with rate $-2\theta/\sigma^2$, and has a mean as follows,

$$\mathsf{E}Z(\infty) = \frac{\sigma^2}{-2\theta},$$

where $Z(\infty)$ stands for a random variable following the stationary distribution of $Z(t)$.

The following lemma describes the least element characterization of the reflection mapping and a relaxed dynamic complementarity property, and is adapted from Section 2 in Chen and Shanthikumar [14].

Lemma 12 (a) (Least element property) Suppose that $x(t)$ is an RCLL function on $[0, \infty)$. If a pair of functions $(y(t), z(t))$ satisfy the following conditions for all $t \geq 0$,

$$z(t) = x(t) + y(t) \geq 0,$$

$$y(t) \text{ is non-decreasing in } t, \text{ with } y(0) = 0,$$

then the following inequalities hold,

$$z(t) \geq \Phi(x)(t) \text{ and } y(t) \geq \Psi(x)(t), \quad \text{for all } t \geq 0.$$

(b) (Relaxed dynamic complementarity property) Suppose that $x(t)$ is an RCLL function on $[0,\infty)$. For any fixed $\epsilon > 0$, if a pair of functions $(y(t), z(t))$ satisfy the following condition for all $t \geq 0$,

$$z(t) = x(t) + y(t) \geq 0,$$
$$y(t) \text{ does not increase at } t \text{ if } z(t) > \epsilon \text{ (or, } (z(t) - \epsilon)dy(t) \leq 0),$$

then, the following inequalities hold,

$$y(t) \leq \Psi(x(\cdot) - \epsilon)(t) \quad \text{and} \quad z(t) - \epsilon \leq \Phi(x(\cdot) - \epsilon)(t).$$

Next, the following strong law of large numbers is concerned with the fluid scaling of the primitive renewal processes defined in (13). It is often used in heavy-traffic analysis. Its proof can be found in the Appendix A.2 of Stolyar [52], which is based on the weak law estimate in Bramson [9].

Lemma 13 Let $t^* > 0$ and $u^* > 0$ be any given time lengths, and assume the condition (12). Then, the following convergence of the fluid scaling given in (13) holds with probability one: as $n \to \infty$,

$$\sup_{0\leq t\leq nt^*} \sup_{0\leq u\leq u^*} |(\bar{E}^n(t+u) - \bar{E}^n(t)) - \lambda u| \to 0,$$
$$\sup_{0\leq t\leq nt^*} \sup_{0\leq u\leq u^*} |(\bar{S}^n(t+u) - \bar{S}^n(t)) - \mu u| \to 0,$$
$$\sup_{0\leq t\leq nt^*} \sup_{0\leq u\leq u^*} |(\bar{\Upsilon}^n(t+u) - \bar{\Upsilon}^n(t)) - \lambda^{-1} u| \to 0.$$

A.2 Proofs for Sections 3 and 4

Proof of Proposition 1. First, we establish that

$$\tilde{B}^n_k(t) \to t, \qquad \text{u.o.c. of } t \geq 0. \tag{129}$$

Write (6-10) for the n-th system with fluid scaling as

$$\tilde{Q}^n_k(t) = \tilde{Q}^n_k(0) + \tilde{\Phi}^n_k(\tilde{E}^n(t)) - \tilde{S}^n_k(\tilde{B}^n_k(t)) \geq 0, \tag{130}$$
$$\tilde{Y}^n_k(t) := \mu_k(t - \tilde{B}^n_k(t)) \text{ is non-decreasing in } t \geq 0, \text{ and } \tilde{Y}^n_k(0) = 0, \tag{131}$$
$$\int_0^\infty \tilde{Q}^n_k(s) d\tilde{Y}^n_k(s) ds = 0. \tag{132}$$

From the assumption in (28), along with the convergence of $\tilde{E}^n(t)$ in (16) and the definition in (18), we can apply random time change theorem (e.g., Chapter 5 of [15]) to obtain

$$\tilde{\Phi}^n_k(\tilde{E}^n(t)) \to p_k \lambda t, \quad \text{u.o.c.} \tag{133}$$

Let $\mathcal{N}_1$ be any subsequence of $\mathcal{N}$. As $\tilde{B}^n_k(t)$ is continuous and non-decreasing, we can find a further subsequence $\mathcal{N}_2$ of $\mathcal{N}_1$, such that as $n \to \infty$ along $\mathcal{N}_2$

$$\tilde{B}^n_k(t) \to \bar{B}_k(t), \quad \text{u.o.c.}, \tag{134}$$

where the limit $\bar{B}_k(t)$ is also (Lipschitz) continuous and nondecreasing. Moreover, by using the convergence of $\tilde{S}^n(t)$ in (16), we have, as $n \to \infty$ along $\mathcal{N}_2$,

$$\tilde{S}^n_k(\tilde{B}^n_k(t)) \to \mu_k \bar{B}_k(t), \quad \text{u.o.c.} \tag{135}$$

Given (30,133,134,135), letting $n \to \infty$ along $\mathcal{N}_2$ in (130-132) yields

$$(\tilde{Q}_k^n(t), \tilde{Y}_k^n(t)) \to (\bar{Q}_k(t), \bar{Y}_k(t)), \quad \text{u.o.c.},$$

where the limit satisfies the followings,

$$\bar{Q}_k(t) = \bar{Y}_k(t) \ [= \mu_k(t - \bar{B}_k(t))] \ \geq 0, \tag{136}$$

$$\bar{Y}_k(t) \text{ is nondecreasing, with } \bar{Y}_k(0) = 0, \tag{137}$$

$$\int_0^\infty \bar{Q}_k(t) d\bar{Y}_k(t). \tag{138}$$

According to the uniqueness of the solution to the one-dimensional Skorohod mapping (cf. Lemma 10), we have

$$\bar{Q}_k(t) = 0 \, (= \Phi(0)) \,, \quad \text{for all } t \geq 0,$$

which implies

$$\bar{B}_k(t) = t, \quad \text{for all } t \geq 0. \tag{139}$$

Since $\mathcal{N}_1$ is arbitrarily given, the convergence in (134) is along the full sequence of $\mathcal{N}$, with $\bar{B}_k(t) = t$, too. That is, the claim in (129) holds.

Next, it follows from (17,28,129), along with the random time-change theorem (e.g., Chapter 5 of [15]), the process $\hat{X}^n(t)$ given in (25) converges weakly as follows,

$$\hat{X}^n(t) \quad \Rightarrow \quad \hat{X}(t), \qquad \text{as } n \to \infty, \tag{140}$$

where the limit $\hat{X}(t)$ is given in (35) and is a Brownian motion with drift θ and a finite covariation matrix. Now, applying the continuity property of the Skorohod mapping (cf. (127,128)) to the Skorohod problem in (22-24), we have the convergence in (31), with the limit satisfying (32-34).

From (32,33), we have

$$\hat{Q}_{\mathcal{K}}(t) = \hat{Q}_{\mathcal{K}}(0) + \hat{X}_{\mathcal{K}}(t) + \hat{Y}_{\mathcal{K}}(t) \geq 0,$$

$$\hat{Y}_{\mathcal{K}}(t) \left(= \sum_{k \in \mathcal{K}} \hat{Y}_k(t) \right) \text{ is non-decreasing in } t \text{ with } \hat{Y}_{\mathcal{K}}(0) = 0.$$

Note that the complementarity property is not necessarily satisfied for $(\hat{Q}_{\mathcal{K}}(t), \hat{Y}_{\mathcal{K}}(t))$. Hence, according to the minimality of the reflection mapping (cf. Lemma 12) we have the lower bound in (36). □

Proof of Proposition 2 (for part (b) only). It suffices to find p_k^n such that the properties in (27,28) hold. Along with (45), this requires that as $n \to \infty$,

$$\frac{\lambda^n p_k^n - \mu_k}{\lambda^n - \mu_{\mathcal{K}}} \to \frac{1}{K} \quad \text{or} \quad \frac{1 - \rho_k^n}{1 - \rho^n} \to \frac{\mu_{\mathcal{K}}}{K \mu_k}, \tag{141}$$

which is indeed (46). Now, given (46), we can use the expression of $\hat{X}_k^n(t)$ in (25) and its convergence in (42) to verify that the property in (28) hold with $\hat{\Phi}_k(t)$ given in (44). □

Proof of Theorem 3. From (2), we have $|\Phi_k^n(E^n(t)) - p_k^n E^n(t)| \leq \kappa$, or, $|\hat{\Phi}_k^n(\tilde{E}^n(t))| \leq \frac{\kappa}{n}$. This implies,

$$\hat{\Phi}_k^n(\tilde{E}^n(t)) \Rightarrow 0, \text{ as } n \to \infty. \tag{142}$$

Similar to the argument for (129), we can show that $\tilde{B}^n(t) \to \mu t$ (u.o.c.) under the given RR policy. Then, for the service process, we also have,

$$\hat{S}_k^n(\tilde{B}_k^n(t)) \Rightarrow \hat{S}_k(t), \text{ as } n \to \infty. \tag{143}$$

Putting the convergences in (142,143,16) and the expression in (25) together yields the convergence of the "free process": $\hat{X}_k^n(t) \Rightarrow \hat{X}_k(t)$ as $n \to \infty$, where $\hat{X}_k(t)$ is given in (51). The diffusion limit for the queue length process is then,

$$\hat{Q}_k^n(t) \Rightarrow \hat{Q}_k(t) := \hat{Q}_k(0) + \hat{X}_k(t) + \hat{Y}_k(t), \tag{144}$$

which satisfies the Skorohod problem (32-34) too and therefore is an RBM with drift θ_k and variation $p_k^2 \lambda c_a^2 + \mu_k c_{b,k}^2$. This RBM has an expected stationary mean as given in (52)

The property in (b) follows directly from the definition of the RR policy and the weak convergence in (a), and the property in (c) can be seen from the discussion prior to the theorem. □

A.3 Proof of Theorem 5: AC Policy

Write $\alpha_k^n(\ell)$ (given in (71)) under the diffusion scaling as, for any (real number) $t \geq 0$,

$$\hat{\alpha}_k^n(t) := \frac{1}{n}\alpha_k^n(\lfloor n^2 t \rfloor) = \hat{\Phi}_k^n(t) - \frac{1}{n}\phi_k^n(\lfloor n^2 t \rfloor) + h_k^n \lambda^n \hat{\Upsilon}^n(t). \tag{145}$$

We adopt the hydrodynamic approach of Bramson [9] (also see, e.g., [43,52,64]) to rewrite $\hat{\alpha}_k^n(t)$. Let $T > 0$ be a (given) constant, and denote for any integer $j \geq 0$ and any $u \geq 0$,

$$\begin{aligned} & \bar{\alpha}_k^{n,j}(u) := \hat{\alpha}_k^n((jT+u)/n) \\ = \; & \frac{1}{n}\Big[\Phi_k^n(\lfloor n(jT+u)\rfloor) - p_k^n \lfloor n(jT+u)\rfloor - \phi_k^n(\lfloor n(jT+u)\rfloor)\Big] \\ & -\frac{h_k^n}{n}\Big[\lfloor n(jT+u)\rfloor - \lambda^n \Upsilon^n(\lfloor n(jT+u)\rfloor)\Big]. \end{aligned}$$

It can be further expressed as,

$$\begin{aligned} & \bar{\alpha}_k^{n,j}(u) = \bar{\alpha}_k^{n,j}(0) + [\bar{\alpha}_k^{n,j}(u) - \bar{\alpha}_k^{n,j}(0)] \\ = \; & \bar{\alpha}_k^{n,j}(0) + \frac{1}{n}\Big[\Phi_k^n(\lfloor n(jT+u)\rfloor) - \Phi_k^n(\lfloor njT\rfloor)\Big] \\ & -\frac{p_k^n}{n}\Big[\lfloor n(jT+u)\rfloor - \lfloor njT\rfloor)\rfloor\Big] - \frac{1}{n}\Big[\phi_k^n(\lfloor n(jT+u)\rfloor) - \phi_k^n(\lfloor njT\rfloor)\rfloor)\Big] \\ & -\frac{h_k^n}{n}\Big[(\lfloor n(jT+u)\rfloor - \lfloor njT\rfloor) - \lambda^n(\Upsilon^n(\lfloor n(jT+u)\rfloor) - \Upsilon^n(\lfloor njT\rfloor))\Big]. \end{aligned} \tag{146}$$

Lemma 14 Let $\Delta > 0$ be any given time length. Consider any sequence of indices $\{j_n\}_{n\in\mathcal{N}}$ satisfying $0 \leq j_n \leq n\Delta/T$, and suppose $\{\bar{\alpha}^{n,j_n}(0)\}_{n\in\mathcal{N}}$ is a bounded (vector) sequence. Then, for any subsequence of $\mathcal{N}$, there exists a further subsequence $\mathcal{N}_1$ such that the followings hold:
(a) (Fluid limit) $\bar{\alpha}^{n,j_n}(u)$ converge to a "fluid limit" $\bar{\alpha}(u)$ as $n \to \infty$ along $\mathcal{N}_1$,

$$\bar{\alpha}_k^{n,j_n}(u) \to \bar{\alpha}_k(u) := \bar{\alpha}_k(0) + \bar{\Phi}_k(u) - p_k u, \quad \text{u.o.c. of } u \geq 0,$$

where $\bar{\Phi}_k(u)$ is a Lipschitz continuous function in u, with a Lipschitz constant 1 and $\bar{\Phi}_k(0) = 0$. Consequently, $\bar{\Phi}_k(u)$ and $\bar{\alpha}_k(u)$ are differentiable for almost all $u \geq 0$.
(b) Denote $\bar{\alpha}_{\min}(u) = \min_{k\in\mathcal{K}} \bar{\alpha}_k(u)$. For any regular time $u \geq 0$ (at which $\bar{\alpha}(u)$ is differentiable), if $\bar{\alpha}_{\min}(u) < 0$, then

$$\dot{\bar{\alpha}}_{\min}(u) \geq \sigma_p := \frac{\min_k p_k}{K}.$$

(c) $\bar{\alpha}(u) = 0$ for all $u \geq |\bar{\alpha}(0)|/\sigma_p$.

Proof. Part (a). Denote for convenience the four terms with the squared brackets in (146) as $f_1^n(u), \cdots, f_4^n(u)$, all with j being replaced by j_n. First, the term $f_2^n(u)$ converges (u.o.c.) to $-p_k u$ as $n \to \infty$. Since $\phi_k^n(\ell)$ is equal to either 0 or 1, the term $f_3^n(u)$ vanishes. According to Lemma 13, the term $f_4^n(u)$ converges (u.o.c.) to 0.

Next, note that $f_1^n(u)$ is nondecreasing in u and satisfies $f_1^n(0) = 0$. Therefore, for any subsequence of $\mathcal{N}$ we can find a further subsequence $\mathcal{N}_1$ such that for some nondecreasing function $\bar{\Phi}_k(u)$ with $\bar{\Phi}_k(0) = 0$, the following convergence holds for all those times $u \geq 0$ where $\bar{\Phi}_k(u)$ is continuous,

$$f_1^n(u) \to \bar{\Phi}_k(u), \quad \text{as } n \to \infty \text{ along } \mathcal{N}_1. \tag{147}$$

Since $\{\bar{\alpha}^{n,j_n}(0)\}$ is a bounded, the subsequence $\mathcal{N}_1$ can be chosen such that, along which, $\bar{\alpha}^{n,j_n}(0) \to \bar{\alpha}(0)$. Note that for any times $0 \leq u_1 \leq u_2$ $(\leq T)$,

$$0 \leq f_1^n(u_2) - f_1^n(u_1) \leq \frac{1}{n}(\lfloor n(jT + u_2) \rfloor - \lfloor n(jT + u_1) \rfloor) \to u_2 - u_1,$$

which implies the Lipschitz continuity of $\bar{\Phi}_k(u)$, with a Lipschitz constant 1 and $\bar{\Phi}_k(0) = 0$. The Lipschitz continuity also implies that the convergence in (147) is not only pointwise, but also u.o.c. of $u \geq 0$.

Part (b). From (75) ($\sum_k \alpha_k^n(\ell) = -1$), we have $\sum_k \bar{\alpha}_k(u) = 0$; hence for any $u \geq 0$, either $\bar{\alpha}(u) = 0$ or $\bar{\alpha}_k(u) < 0$ for some $k \in \mathcal{K}$. Consider any regular time $u_0 \geq 0$ such that $\bar{\alpha}_{\min}(u_0) < 0$. Denote $\mathcal{K}^{\min} = \arg\min_{k \in \mathcal{K}} \bar{\alpha}_k(u_0)$, i.e., the set of servers that have minimum (and negative) surplus of routing deviation (cf. the definition in (71-73) and the remark following the definition). From the conclusion (a), we can find (small) positive constants δ and ϵ such that for any $k_1 \in \mathcal{K}^{\min}$, $k_2 \in \mathcal{K} \setminus \mathcal{K}^{\min}$, $u \in [u_0, u_0 + \delta]$, we have the following for sufficiently large $n \in \mathcal{N}_1$,

$$\bar{\alpha}_{k_2}^{n,j_n}(u) - \bar{\alpha}_{k_1}^{n,j_n}(u) > \epsilon.$$

This inequality implies that $\mathcal{K}^{\min}$ must contain the server with the minimum surplus of routing deviation during the time interval $[u_0, u_0 + \delta]$ for the fluid scaled process $\bar{\alpha}^{n,j_n}(u)$ (i.e., the ℓ-th job, $\lfloor n(j_n T + u_0) \rfloor \leq \ell \leq \lfloor n(j_n T + u_0 + \delta) \rfloor$ in the original scale of $\alpha^{n,j_n}(\ell)$) for sufficiently large $n \in \mathcal{N}_1$. Hence, according to the policy in (71,72), all jobs are routed to servers in $\mathcal{K}^{\min}$ during this time interval; that is, for any time $u \in [u_0, u_0 + \delta]$,

$$\sum_{k \in \mathcal{K}^{\min}} [\Phi_k^n(\lfloor n(j_n T + u) \rfloor) - \Phi_k^n(\lfloor n(j_n T + u_0) \rfloor)] = \lfloor n(j_n T + u) \rfloor - \lfloor n(j_n T + u_0) \rfloor. \tag{148}$$

Applying the above to (146) (for $j = j_n$), with simple algebra, we have the following,

$$\begin{aligned}
& \sum_{k \in \mathcal{K}^{\min}} \left[\bar{\alpha}_k^{n,j_n}(u) - \bar{\alpha}_k^{n,j_n}(u_0) \right] \\
= \; & \frac{1}{n} \left(1 - \sum_{k \in \mathcal{K}^{\min}} p_k^n \right) \left[\lfloor n(j_n T + u) \rfloor - \lfloor n(j_n T + u_0)) \rfloor \right] \\
& - \frac{1}{n} \sum_{k \in \mathcal{K}^{\min}} \left[\phi_k^n(\lfloor n(j_n T + u) \rfloor) - \phi_k^n(\lfloor n(j_n T + u_0)) \rfloor) \right] \\
& - \sum_{k \in \mathcal{K}^{\min}} \frac{h_k^n}{n} \Big[(\lfloor n(j_n T + u) \rfloor - \lfloor n(j_n T + u_0) \rfloor) \\
& - \lambda^n(\Upsilon^n(\lfloor n(j_n T + u) \rfloor) - \Upsilon^n(\lfloor n(j_n T + u_0) \rfloor)) \Big].
\end{aligned} \tag{149}$$

Letting $n \to \infty$ along $\mathcal{N}_1$, we have

$$\sum_{k \in \mathcal{K}^{\min}} [\bar{\alpha}_k(u) - \bar{\alpha}_k(u_0)] = (1 - \sum_{k \in \mathcal{K}^{\min}} p_k)(u - u_0).$$

Since the above equality holds for all $u \in [u_0, u_0 + \delta]$ and the time u_0 is regular, the above implies

$$\sum_{k \in \mathcal{K}^{\min}} \dot{\bar{\alpha}}_k(u_0) = 1 - \sum_{k \in \mathcal{K}^{\min}} p_k.$$

Observe that

$$\dot{\bar{\alpha}}_k(u_0) = \dot{\bar{\alpha}}_{\min}(u_0) \quad \text{for } k \in \mathcal{K}^{\min}. \tag{150}$$

(Refer to the proof of Proposition 3(c) of [16] and Lemma 3.2 of [20] for similar cases.) Therefore, we have

$$\dot{\bar{\alpha}}_{\min}(u_0) = \frac{1}{|\mathcal{K}^{\min}|}\Big(1 - \sum_{k \in \mathcal{K}^{\min}} p_k\Big) \geq \sigma_p.$$

The conclusion in (c) follows from the part (b) immediately. □

Lemma 15 Let $\Delta > 0$ be any given time and $\epsilon > 0$ be any (small) number. Then, there exists a sufficiently large time T such that for sufficiently large n, the following holds for all integer $0 \leq j \leq n\Delta/T$ and $u \in [0, T]$,

$$|\bar{\alpha}^{n,j}(u)| < \epsilon. \tag{151}$$

Consequently, we have

$$\hat{\alpha}^n(t) \to 0 \quad \text{u.o.c. of } t \geq 0, \text{ as } n \to \infty. \tag{152}$$

Proof. Choose any time length T such that $T \geq \epsilon/\sigma_p$, where σ_p is specified in Lemma 14. This time length T is long enough so that the fluid limit $\bar{\alpha}(t)$ will reach zero, from any initial point $|\bar{\alpha}(0)| \leq \epsilon$.

We prove the property for $j = 0$ first. Note that by way of the construction, we have $\bar{\alpha}^{n,0}(0) = \hat{\alpha}^n(0) = 0$, and hence, from Lemma 14(a,c), we have, as $n \to \infty$ (here, along the full sequence $\mathcal{N}$),

$$\bar{\alpha}^{n,0}(u) \to 0 \qquad \text{u.o.c.} \tag{153}$$

The above implies the conclusion in (151) for $j = 0$, for sufficiently large n.

Next, we extend to verify the property in (151) for $j = 1, \ldots, n\Delta/T$. Suppose to the contrary, there exists a subsequence $\mathcal{N}_1$ of $\mathcal{N}$ such that, for any $n \in \mathcal{N}_1$, the property in (151) does not hold for some integers $1 \leq j \leq n\delta/T$. Consequently, for any $n \in \mathcal{N}_1$, there exists a smallest integer, denoted as j_n, in the interval $[1, n\Delta/T]$ such that the property in (151) does not hold. To reach a contradiction, it suffices to construct an infinite subsequence $\mathcal{N}_2 \subset \mathcal{N}_1$, such that the desired property in (151) hold for $j = j_n$ for sufficiently large $n \in \mathcal{N}_2$.

From the proof of the property in (151) for $j = 0$ and the contradictory assumption above, we know that the property (151) holds for $j = 0, ..., j_n - 1$, $n \in \mathcal{N}_1$. Specifically, for $j = j_n - 1$, we have

$$|\bar{\alpha}^{n,j_n-1}(0)| \leq \epsilon, \quad \text{for all (sufficiently large) } n \in \mathcal{N}_1.$$

Therefore, the sequence $\{\bar{\alpha}^{n,j_n-1}(0), n \in \mathcal{N}_1\}$ has a convergent subsequence. Then, by Lemma 14(a,c) again, there exists a further subsequence $\mathcal{N}_2 \subset \mathcal{N}_1$ such that

$$\bar{\alpha}^{n,j_n-1}(u) \to \bar{\alpha}(u), \qquad \text{u.o.c., as } n \to \infty \text{ along } \mathcal{N}_2, \tag{154}$$

with $|\bar{\alpha}(0)| \leq \epsilon$ and $\bar{\alpha}(u) = 0$ for $u \geq T(\geq \epsilon/\sigma_p)$. Recall from the definition that $\bar{\alpha}^{n,j_n}(u) = \bar{\alpha}^{n,j_n-1}(T+u)$. Hence, from (154), we have

$$\bar{\alpha}^{n,j_n}(u) \to 0, \qquad \text{u.o.c., as } n \to \infty \text{ along } \mathcal{N}_2,$$

which implies the property in (151) for $j = j_n$ for sufficiently large $n \in \mathcal{N}_2$.

Following the definitions of $\hat{\alpha}^n(t)$ and $\bar{\alpha}^{n,j}(u)$, the first conclusion of the lemma translates to the following immediately: For any $\Delta > 0$ and $\epsilon > 0$, we have for sufficiently large n,

$$|\hat{\alpha}^n(t)| < \epsilon, \quad t \in [0, \Delta].$$

This implies the convergence in (152). □

Proof (of Theorem 5). Applying Lemma 15 and the functional central limit theorem for $\hat{\Upsilon}^n(t)$ (refer to (17)) to the expression in (145), we have the following weak convergence,

$$\hat{\Phi}_k^n(\lambda^n t) \Rightarrow \hat{\Phi}_k(\lambda t) := h_k \hat{E}(t).$$

Given this convergence, it is direct to verify the conclusion in (b). And then, the conclusion in (a) follows from Proposition 1. Finally, for the conclusion in (c), it can be verified directly that the parameters (p^n, h^n) specified in (80) meets the convergence requirements in (27,77) with the limits (θ^*, h^*) following the relationship in (67). The other conclusions in (c) can be seen from the discussions leading to the lower bound in (68).

A.4 Proof of Theorem 6: SC Policy

Write $\beta_k^n(\ell)$ (given in (91)) under the diffusion scaling as, for any $t \geq 0$,

$$\hat{\beta}_k^n(t) := \frac{1}{n}\beta_k^n(\lfloor n^2 t \rfloor) = \hat{\Phi}_k^n(t) - \frac{1}{n}\phi_k^n(\lfloor n^2 t \rfloor) - \Big[\hat{S}_k^n(\tilde{B}_k^n(\tilde{\Upsilon}^n(t)) - \sum_{i \in \mathcal{K}} h_{ki}^n \hat{S}_i^n(\tilde{B}_i^n(\tilde{\Upsilon}^n(t)))\Big]. \quad (155)$$

Let $T > 0$ be a (given) constant, and denote for any integer $j \geq 0$ and any $u \geq 0$,

$$\bar{\beta}_k^{n,j}(u) := \hat{\beta}_k^n((jT+u)/n) = \frac{1}{n}\beta_k^n(\lfloor njT + nu \rfloor).$$

It can be expressed as,

$$\begin{aligned} \bar{\beta}_k^{n,j}(u) &= \bar{\beta}_k^{n,j}(0) + \frac{1}{n}\Big[\Phi_k^n(\lfloor n(jT+u) \rfloor) - \Phi_k^n(\lfloor njT \rfloor)\Big] - \frac{p_k^n}{n}\Big[\lfloor njT + nu \rfloor - \lfloor njT \rfloor\Big] \\ &\quad -\frac{1}{n}\Big[\phi_k^n(\lfloor n(jT+u) \rfloor) - \phi_k^n(\lfloor njT \rfloor)\Big] - \Big[\eta_k^{n,j}(u) - \sum_{i \in \mathcal{K}} h_{ki}^n \eta_i^{n,j}(u)\Big], \end{aligned} \quad (156)$$

where

$$\begin{aligned} \eta_k^{n,j}(u) &= \frac{1}{n}\Big[S_k^n(B_k^n(\Upsilon^n(\lfloor njT + nu \rfloor))) - S_k^n(B_k^n(\Upsilon^n(\lfloor njT \rfloor)))\Big] \\ &\quad -\frac{1}{n}\Big[\mu_k B_k^n(\Upsilon^n(\lfloor njT + nu \rfloor)) - \mu_k B_k^n(\Upsilon^n(\lfloor njT \rfloor))\Big]. \end{aligned}$$

Lemma 16 Let $\Delta > 0$ be any given time length, and consider any sequence of indices $\{j_n\}_{n \in \mathcal{N}}$ satisfying $0 \leq j_n \leq n\Delta/T$. Then, the following holds (almost surely),

$$\eta_k^{n,j_n}(u) \to 0, \quad \text{u.o.c. of } u \geq 0. \quad (157)$$

Proof. Write

$$\eta_k^{n,j}(u) = [\bar{S}_k^n(t^n + s^n) - \bar{S}_k^n(t^n)] - \mu_k s^n, \quad \text{with}$$
$$t^n = \bar{B}_k^n(\bar{\Upsilon}^n(j_n T)), \quad s^n = \bar{B}_k^n(\bar{\Upsilon}^n(j_n T + u)) - \bar{B}_k^n(\bar{\Upsilon}^n(j_n T)),$$

where $(\bar{B}_k^n(t), \bar{\Upsilon}_k^n(t)) = \frac{1}{n}(B_k^n(nt), \Upsilon_k^n(\lfloor nt \rfloor))$ (as in (13,14)). Estimate

$$\frac{1}{n}t^n \quad = \quad \frac{1}{n}\bar{B}_k^n(\bar{\Upsilon}^n(j_n T)) \leq \frac{1}{n}\bar{\Upsilon}^n(j_n T) \leq \frac{1}{n}\bar{\Upsilon}^n(n\Delta) \to \frac{\Delta}{\lambda}.$$

Hence, for sufficiently large n, we have $t^n \leq nt^*$ with $t^* := \lambda^{-1}\Delta + 1$. Moreover, we have the following estimate,

$$s^n \leq \bar{\Upsilon}^n(j_n T + u) - \bar{\Upsilon}^n(j_n T) \leq |\bar{\Upsilon}^n(j_n T + u) - \bar{\Upsilon}^n(j_n T) - \lambda^{-1}u| + \lambda^{-1}u \to \lambda^{-1}u,$$

where the first inequality is due to the equation in (7) and the convergence follows from Lemma 13. Hence, letting $\tau > 0$ be an arbitrarily given time length, we have for sufficiently large n, $s^n \leq s^* := \lambda^{-1}\tau + 1$ for all $u \leq \tau$.

Given the above estimates of t^n and s^n, we have for sufficiently large n, for all $u \in [0, \tau]$,

$$\eta_k^{n,j_n}(u) \quad \leq \quad \sup_{0 \leq t \leq nt^*} \sup_{0 \leq s \leq s^*} |\bar{S}^n(t+s) - \bar{S}^n(t) - \mu_k s|.$$

Then, applying Lemma 13, we have the convergence in (157). □

Lemma 17 Let $\Delta > 0$ be any given time length. Consider any sequence of indices $\{j_n\}_{n \in \mathcal{N}}$ satisfying $0 \leq j_n \leq n\Delta/T$, and suppose $\{\bar{\beta}^{n,j_n}(0)\}_{n \in \mathcal{N}}$ is a bounded (vector) sequence. Then, for any subsequence of $\mathcal{N}$, there exists a further subsequence $\mathcal{N}_1$ such that the followings hold:
(a) (Fluid limit) $\bar{\beta}^{n,j_n}(u)$ converge to a "fluid limit" $\bar{\beta}(u)$ as $n \to \infty$ along $\mathcal{N}_1$,

$$\bar{\beta}_k^{n,j_n}(u) \to \bar{\beta}_k(u) := \bar{\beta}_k(0) + \bar{\Phi}_k(u) - p_k u, \quad \text{u.o.c. of } u \geq 0,$$

where $\bar{\Phi}_k(u)$ is a Lipschitz continuous function in u, with a Lipschitz constant 1 and $\bar{\Phi}_k(0) = 0$. Consequently, $\bar{\Phi}_k(u)$ and $\bar{\beta}_k(u)$ are differentiable for almost all $u \geq 0$.
(b) Denote $\bar{\beta}_{\min}(u) = \min_{k \in \mathcal{K}} \bar{\beta}_k(u)$. For any regular time $u \geq 0$ (at which $\bar{\beta}(u)$ is differentiable), if $\bar{\beta}_{\min}(u) < 0$, then

$$\dot{\bar{\beta}}_{\min}(u) \geq \sigma_p = \frac{\min_k p_k}{K}.$$

(c) $\bar{\beta}(u) = 0$ for all $u \geq |\bar{\beta}(0)|/\sigma_p$.

Proof. According to Lemma 16, the term in the last squared bracket in (156) converge to 0. Then, the rest of the proof for the conclusion (a) simply repeats the one for Lemma 14(a).

The proof for the conclusion (b) involves two slight modifications of the corresponding one for Lemma 14(b). First, the equality in (148) is now argued by using the SC policy specified in (90,92). Second, the equation (149) is modified as follows,

$$\begin{aligned}
&\sum_{k \in \mathcal{K}^{\min}} [\bar{\beta}_k^{n,j_n}(u) - \bar{\beta}_k^{n,j_n}(u_0)] \\
= \quad & \frac{1}{n}(1 - \sum_{k \in \mathcal{K}^{\min}} p_k^n)\,[\lfloor n(jT+u) \rfloor - \lfloor n(jT+u_0)) \rfloor] \\
& - \frac{1}{n} \sum_{k \in \mathcal{K}^{\min}} [\phi_k^n(\lfloor n(jT+u) \rfloor) - \phi_k^n(\lfloor n(jT+u_0)) \rfloor)] \\
& - \sum_{k \in \mathcal{K}^{\min}} \left[(\eta_k^{n,j_n}(u) - \eta_k^{n,j_n}(u_0)) - \sum_{i \in \mathcal{K}} h_{ki}^n (\eta_i^{n,j_n}(u) - \eta_i^{n,j_n}(u_0)) \right].
\end{aligned}$$

The conclusion in (c) also follows from part (b) immediately. □

Lemma 18 Let $\Delta > 0$ be any given time and $\epsilon > 0$ be any (small) number. Then, there exists a sufficiently large time T such that for sufficiently large n, the following holds for all integer $0 \leq j \leq n\Delta/T$ and $u \in [0, T]$,

$$|\bar{\beta}^{n,j}(u)| < \epsilon.$$

Consequently, we have

$$\hat{\beta}^n(t) \to 0 \ \text{ u.o.c. of } t \geq 0, \text{ as } n \to \infty. \tag{158}$$

The proof of this lemma is a repetition of the one for Lemma 15, where the role of Lemma 14 will be replaced by Lemma 17 just established. Hence, the detailed proof is omitted.

Proof (of Theorem 6). Write

$$\begin{aligned}\tilde{\beta}^n_k(t) = \frac{1}{n}\hat{\beta}^n_k(t) &= \tilde{\Phi}^n_k(t) - \frac{1}{n^2}p^n_k\lfloor n^2 t\rfloor - \frac{1}{n^2}\phi^n_k(\lfloor n^2 t\rfloor) \\ &\quad -\frac{1}{n}\left[\hat{S}^n_k(\tilde{B}^n_k(\tilde{\Upsilon}^n(t)) - \sum_{i\in\mathcal{K}} h^n_{ki}\hat{S}^n_i(\tilde{B}^n_i(\tilde{\Upsilon}^n(t)))\right].\end{aligned}$$

From Lemma 18, we have $\tilde{\beta}^n_k(t) \to 0$ (u.o.c.) as $n \to \infty$. Note that the terms involving ϕ^n_k and $\hat{S}^n$ in the above all vanish as $n \to \infty$ too. Therefore, letting $n \to \infty$ in the above yields

$$\tilde{\Phi}^n_k(t) \to p_k t, \ \text{ u.o.c. of } t \geq 0, \text{ as } n \to \infty.$$

Using the this convergence, we can repeat the argument from (133) to (139) to show the following,

$$\tilde{B}^n_k(t) \to \bar{B}_k(t) \equiv t, \ \text{ u.o.c. of } t \geq 0, \text{ as } n \to \infty. \tag{159}$$

Recall that

$$\tilde{\Upsilon}^n(\lambda^n t) \to t \ \text{ and } \ \hat{S}^n(t) \Rightarrow \hat{S}(t). \tag{160}$$

Now, applying the convergences in (158,159,160) to the definition in (155), we have

$$\hat{\Phi}^n_k(\lambda^n t) \Rightarrow \hat{\Phi}_k(\lambda t) := \hat{S}_k(t) + \sum_{i\in\mathcal{K}} h_{ki}\hat{S}_i(t). \tag{161}$$

Similar to the proof of Theorem 1, given the above convergence, it is direct to verify the conclusion in (a,b).

For the conclusion in (c), it can be verified directly that the parameters (p^n, h^n) specified in (100) meet the convergence requirements in (27,96) with the limit $(\theta, h) = (\theta^*, h^*)$ following the relationship in (86). The optimality property in (101) follows from the discussions leading to the lower bound in (87). The equality (102) follows from (161,35).

From (102) and the Skorohod mapping (Lemma 10), we have

$$\hat{Q}_k(t) = \hat{Q}_k(0) + p_k\hat{X}_{\mathcal{K}}(t) + \sup_{0\leq s\leq t}(-\hat{Q}_k(0) - p_k\hat{X}_{\mathcal{K}}(s))^+.$$

Note that the time τ, given in (104), is the first time that the "free" process $(-\hat{X}_{\mathcal{K}}(s))$ increases to the maximum initial queue length (weighted by p_k) among all classes. Hence, it can be verified that at time τ,

$$\sup_{0\leq s\leq \tau}(-\hat{Q}_k(0) - p_k\hat{X}_{\mathcal{K}}(s))^+ = -\hat{Q}_k(0) - p_k\hat{X}_{\mathcal{K}}(\tau),$$

and therefore, $\hat{Q}_k(\tau) = 0$ for all k. Now, we "restart" the system $\hat{Q}(t)$ from the time τ and write

$$\hat{Q}_k(\tau + t) = \hat{Q}_k(\tau) + p_k(\hat{X}_{\mathcal{K}}(\tau + t) - \hat{X}_{\mathcal{K}}(\tau)) + (\hat{Y}_k(\tau + t) - \hat{Y}_k(\tau)),$$

or

$$\frac{1}{p_k}\hat{Q}_k(\tau+t) = (\hat{X}_{\mathcal{K}}(\tau+t) - \hat{X}_{\mathcal{K}}(\tau)) + \frac{1}{p_k}(\hat{Y}_k(\tau+t) - \hat{Y}_k(\tau)).$$

Clearly, the triples $(\frac{1}{p_k}\hat{Q}_k(\tau+\cdot), (\hat{X}_{\mathcal{K}}(\tau+\cdot) - \hat{X}_{\mathcal{K}}(\tau)), \frac{1}{p_k}(\hat{Y}_k(\tau+\cdot) - \hat{Y}_k(\tau))$, $k \in \mathcal{K}$, constitutes a one-dimensional Skorohod mapping. Since these Skorohod mappings (for all k) are driven by the same "free" process $(\hat{X}_{\mathcal{K}}(\tau+\cdot) - \hat{X}_{\mathcal{K}}(\tau))$, all the queue lengths $\frac{1}{p_k}\hat{Q}_k(\tau+\cdot)$ are equal. That is, the conclusion in (103) holds.

The expected stationary queue lengths in (105) also follows from the discussions leading to the lower bound in (87).

A.5 Proof of Theorem 7: Stability and Stationarity under RR, AC, SC, JSQ and BR Policies

We apply the fluid model approach to prove Theorem 7. Lemma 19 below relates the n-th (diffusion-scaled) system $\hat{\Xi}^n(t)$ to the fluid model characterized by the following relationships:

$$\begin{aligned}\hat{q}_k^n(t) &= \hat{q}_k^n(0) + n\bar{\phi}_k^n\left(\lambda^n(t - t\wedge\frac{\hat{u}^n(0)}{n})\right) - n\mu_k\left(\bar{b}_k^n(t) - \bar{b}_k^n(t)\wedge\frac{\hat{v}_k^n(0)}{n}\right)\\ &= \hat{q}_k^n(0) + n\bar{\phi}_k^n\left(\lambda^n(t - t\wedge\frac{\hat{u}^n(0)}{n})\right) - n\mu_k\left(t - \bar{b}_k^n(t)\wedge\frac{\hat{v}_k^n(0)}{n}\right) + \hat{y}_k^n(t) \geq 0, \quad (162)\end{aligned}$$

$$\bar{\phi}_k^n(t) \text{ is non-decreasing with } \bar{\phi}_k^n(0) = 0, \text{ and Lipschitz with constant 1,} \quad (163)$$

$$\bar{b}_k^n(t) \text{ is non-decreasing with } \bar{b}_k^n(0) = 0, \text{ and Lipschitz with constant 1,} \quad (164)$$

$$\hat{y}_k^n(t) = n\mu_k(t - \bar{b}_k^n(t)) \text{ is non-decreasing with } \hat{y}_k^n(0) = 0, \quad (165)$$

$$\int_0^\infty \hat{q}_k^n(s)d\hat{y}_k^n(s) = 0, \quad (166)$$

and the fluid analogy of the chasing process, $\hat{\psi}_k^n(t)$, satisfies the followings under the respective policies,

$$\begin{aligned}\hat{\psi}_k^n(t; AC(\theta,h)) &= \hat{\psi}_k^n(0) + n\bar{\phi}_k^n(\lambda^n(t - t\wedge\frac{\hat{u}^n(0)}{n})) - p_k^n\lambda^n n(t - t\wedge\frac{\hat{u}^n(0)}{n})\\ &\quad + h_k^n\lambda^n n(t\wedge\frac{\hat{u}^n(0)}{n}), \quad (167)\\ \hat{\psi}_k^n(t; SC(\theta,h)) &= \hat{\psi}_k^n(0) + n\bar{\phi}_k^n(\lambda^n(t - t\wedge\frac{\hat{u}^n(0)}{n})) - p_k^n\lambda^n n(t - t\wedge\frac{\hat{u}^n(0)}{n})\\ &\quad + \left[\mu_k n(\bar{b}_k^n(t)\wedge\frac{\hat{v}_k^n(0)}{n}) - \sum_{i\in\mathcal{K}} h_{ki}^n\mu_k n(\bar{b}_i^n(t)\wedge\frac{\hat{v}_i^n(0)}{n})\right], \quad (168)\\ \hat{\psi}_k^n(t; JSQ/BR) &= 0. \quad (169)\end{aligned}$$

Note the "hat" and "bar" designations in the above processes, such as $\hat{q}_k^n(t)$ and $\bar{b}_k^n(t)$, are in line with the scalings of their stochastic counterparts, such as the diffusion-scaled process $\hat{Q}_k^n(t)$ and the fluid-scaled process $\tilde{B}_k^n(t)$.

Lemma 19 Suppose the $RR(\theta)$, $AC(\theta,h)$, $SC(\theta,h)$, JSQ or BR policy is in force, and consider the n-th (diffusion-scaled) system characterized by the Markov process $\Xi^n(t)$, for any fixed n. Let $\{m_i; i = 1,2,\cdots\}$ be a sequence of numbers such that $m_i \to \infty$ as $i \to \infty$; and let $\{x^i \in \mathcal{X}; i = 1,2,\cdots\}$ be a sequence of initial states such that $|x^i| \leq m_i$ for all i.
(a) Then, for any subsequence of positive integers, there exists a further subsequence, denoted by $\mathcal{I}$, such that the following (a.s.) convergence holds as $i \to \infty$ along $\mathcal{I}$,

$$\frac{1}{m_i}\hat{\Xi}^n(0;x^i) = \frac{1}{m_i}\left(\hat{Q}^n(0), \hat{U}^n(0), \hat{V}^n(0), \hat{\Psi}^n(0)\right) \to \left(\hat{q}^n(0), \hat{u}^n(0), \hat{v}^n(0), \hat{\psi}^n(0)\right), \quad (170)$$

and

$$\frac{1}{m_i}\left(\hat{Q}^n(m_it),\tilde{\Phi}^n(m_it),\tilde{B}^n(m_it),\hat{Y}^n(m_it),\hat{\Psi}^n(m_it)\right)$$
$$\to\left(\hat{q}^n(t),\bar{\phi}^n(t),\bar{b}^n(t),\hat{y}^n(t),\hat{\psi}^n(t)\right),\quad \text{u.o.c. of } t\geq 0, \tag{171}$$

where the limit is Lipschitz continuous and is a solution to the fluid model in (162-169), with initial condition

$$|\hat{q}^n(0)|+|\hat{u}^n(0)|+|\hat{v}^n(0)|+|\hat{\psi}^n(0)|\leq 1. \tag{172}$$

(b) (Uniform stability) Moreover, there exists a time $t_0>0$, independent of the index n, such that for sufficiently large n, any fluid limit derived in part (a) (which is a solution to (162-169,172)) have

$$\hat{q}^n(t)=0 \text{ and } \hat{\psi}^n(t)=0,\quad t\geq t_0. \tag{173}$$

Proof. *Part (a), for $AC(\theta,h)$, $SC(\theta,h)$, JSQ and BR policies.* As the initial state is bounded by 1 ($|\hat{\Xi}^n(0;x^i)|/m_i=|x^i|/m_i\leq 1$), it follows that there exists a subsequence $\mathcal{I}$ such that the convergence of initial state in (170) with a limit satisfying (172). Hence, we have, as $i\to\infty$ along $\mathcal{I}$,

$$\frac{1}{m_i}\tilde{E}^n(m_it)\to\lambda^n(t-t\wedge\frac{\hat{u}^n(0)}{n}),\ \frac{1}{m_i}\tilde{S}^n_k(m_it))\to\mu_k(t-t\wedge\frac{\hat{v}^n_k(0)}{n}),\quad \text{u.o.c. of } t\geq 0, \tag{174}$$

by applying the functional strong law-of-large-numbers (e.g., Theorem 5.10 in [15]) and carefully taking care of the initial residuals (e.g., Lemma 4.2 of [17], Section A.1 of [66]). Then, following the procedure from (130) to (138) in the proof of Proposition 1, with the parameter m_i playing the role of scaling factor, it is direct to establish the convergence of the first four components in (171), with the limit satisfying (162-166). Similarly, using the same approach for Lemmas 14(a) and 17(a), we establish the convergence of $\hat{\Psi}^n(m_it)/m_i$ with the limit satisfying (167) and (168) under $AC(\theta,h)$ and $SC(\theta,h)$ policies respectively. (Recall that the RR policy can be viewed as a special case of the AC and SC policies.)

Part (b), for $AC(\theta,h)$ and $SC(\theta,h)$ policies. Observe from (115) that $V^n_k(0)=v^n_{k,1}$. If $V^n_k(0)>0$, then there must be an initial class-k jobs being served ($Q^n_k(0)>0$), and this initial service completes at the time $V^n_k(0)$ (or $v^n_{k,1}$). Thus, we have $B^n_k(t)=t$ for $t<V^n_k(0)$, which implies $\bar{b}^n_k(t)=t$ for $t\leq\hat{v}^n_k(0)/n$. And the term in (162), $\bar{b}^n_k(t)\wedge(\hat{v}^n_k(0)/n)$, then can be replaced by $t\wedge(\hat{v}^n_k(0)/n)$. Consequently, after a time longer than all (scaled) residuals, say, for $t\geq 1/n\geq(|\hat{u}^n(0)|+|\hat{v}^n(0)|)/n$, the relationships in (162,167,168) are reduced to

$$\hat{q}^n_k(t)=\hat{q}^n_k(0)+n\bar{\phi}^n_k(\lambda^n(t-\frac{\hat{u}^n(0)}{n}))-n\mu_k\left(\bar{b}^n_k(t)-\frac{\hat{v}^n_k(0)}{n}\right), \tag{175}$$

$$\hat{\psi}^n_k(t)=\hat{\psi}^n_k(0)+n\bar{\phi}^n_k(\lambda^n(t-\frac{\hat{u}^n(0)}{n}))-p^n_k\lambda^n n(t-\frac{\hat{u}^n(0)}{n})+A^n_k, \tag{176}$$

where for $AC(\theta,h)$,

$$A^n_k=h^n_k\lambda^n\hat{u}^n(0)\leq h^n_k\lambda^n,$$

and for $SC(\theta,h)$,

$$A^n_k=\mu_k\hat{v}^n_k(0)-\sum_{i\in\mathcal{K}}h^n_{ki}\mu_k\hat{v}^n_i(0)\leq\mu_k+\sum_{i\in\mathcal{K}}h^n_{ki}\mu_k.$$

Moreover, at time $t = \frac{1}{n}$, the state can be bounded as,

$$\begin{aligned}
& |\hat{q}^n(\frac{1}{n})| + |\hat{\psi}^n(\frac{1}{n})| \\
\leq & \sum_{k\in\mathcal{K}} \left[\hat{q}_k^n(0) + n\bar{\phi}_k^n(\lambda^n(\frac{1}{n} - \frac{\hat{u}^n(0)}{n})) + n\mu_k \left(\bar{b}_k^n(\frac{1}{n}) - \frac{\hat{v}_k^n(0)}{n} \right) \right] \\
& + \sum_{k\in\mathcal{K}} \left[\hat{\psi}_k^n(0) + n\bar{\phi}_k^n(\lambda^n(\frac{1}{n} - \frac{\hat{u}^n(0)}{n})) + p_k^n\lambda^n n(\frac{1}{n} - \frac{\hat{u}^n(0)}{n}) + A_k^n \right] \\
\leq & |\hat{q}^n(0)| + \sum_{k\in\mathcal{K}} [\lambda^n + \mu_k] + |\hat{\psi}^n(0)| + \sum_{k\in\mathcal{K}} [\lambda^n + p_k^n\lambda^n + A_k^n] \\
\leq & \kappa_1, \qquad (177)
\end{aligned}$$

for some constant $\kappa_1 > 0$ that is independent of n.

Next, using the approach for proving Lemma 14(b) for AC policy (or Lemma 17(b) for SC policy), we can find a time length $\sigma > 0$ that is independent of n and satisfies $\sigma < \sigma_p = \min_k p_k^n/K$ (for sufficiently large n), such that

$$\hat{\psi}^n(t) = 0, \quad \text{for } t \geq \frac{1}{n} + \frac{\hat{\psi}^n(\frac{1}{n})/n}{\sigma}.$$

(Here, the function $\hat{\psi}^n(\frac{1}{n} + t)/n$ plays the role of $\bar{\alpha}(t)$ in Lemma 14, or $\bar{\beta}(t)$ in Lemma 17. In addition, in that lemma the limit is derived by taking $n \to \infty$ while here we scale the (fixed) n-th system by m_i and let $m_i \to \infty$.) Let $\tau = 1 + \kappa_1/\sigma$. Given the bound in (177), we have $\tau \geq 1 + \hat{\psi}^n(\frac{1}{n})/\sigma$. Hence, the above implies

$$\hat{\psi}^n(t) = 0, \quad \text{for } t \geq \frac{\tau}{n}, \qquad (178)$$

and similar to (177), at time $t = \tau/n$, we have

$$|\hat{q}^n(\frac{\tau}{n})| + |\hat{\psi}^n(\frac{\tau}{n})| \leq \kappa_2. \qquad (179)$$

Now, using the property in (178) and taking derivative at both sides of (176), we have for $t \geq \tau/n$,

$$\frac{d}{dt}\bar{\phi}_k^n(\lambda^n(t - \frac{\hat{u}^n(0)}{n})) = p_k^n\lambda^n. \qquad (180)$$

Consider any time $t \geq \tau/n$ and $\hat{q}_k^n(t) > 0$. From (165,166), we have $d\hat{y}_k^n(t)/dt = 0$ and hence $\dot{\bar{b}}_k^n(t) = 1$; that is, in the fluid limit, the server k is busy if there is a positive queue length. Given the latter and by taking derivative on (175), we have

$$\frac{d\hat{q}_k^n(t)}{dt} = np_k^n\lambda^n - n\mu_k = \theta_k^n. \qquad (181)$$

Recall that the right-hand-side of the above has a limit $\theta_k < 0$. Hence, from (179,181), we can find a time t' that is independent of n, such that

$$\hat{q}_k^n(t) = 0, \quad \text{for } t \geq \frac{\tau}{n} + t'.$$

Consequently, the conclusion (a) holds with $t_0 := 1 + t'$.

Part (b), for JSQ and BR policy. We deal with the BR policy (with $2 \leq c \leq K$ and $\pi_k^n = \mu_k/\mu_{\mathcal{K}}$, as in Proposition 2) only, as the JSQ policy is a special case.

First, it can be verified that under the BR policy, the results in (175,177) are still valid for $t \geq 1/n$ (with $\hat{\psi}_k^n(t) \equiv 0$).

Denote $\hat{q}_{\max}^n(t) := \max_{k\in\mathcal{K}} \hat{q}_k^n(t)$. Let $\mathcal{K}_{\max}^t = \operatorname{argmax}_{k\in\mathcal{K}} \hat{q}_k^n(t)$ be the set of the servers with the maximum fluid level and let $K_{\max}^t$ be the cardinal of (the number of elements in) the set $\mathcal{K}_{\max}^t$ at time $t \geq 0$. Define $\hat{q}_{\min}^n(t)$, $\mathcal{K}_{\min}^t$ and $K_{\min}^t$ similarly. Similar to (150), the following result is widely known: for any $k \in \mathcal{K}_{\max}^t$ and $\ell \in \mathcal{K}_{\min}^t$,

$$\dot{\hat{q}}_k^n(t) = \dot{\hat{q}}_{\max}^n(t) \quad \text{and} \quad \dot{\hat{q}}_\ell^n(t) = \dot{\hat{q}}_{\min}^n(t). \tag{182}$$

Also recall the following result from Proposition 3(b) of [16]. If $\max_{k\in\mathcal{K}_1} \hat{q}_k^n(t) < \min_{k\in\mathcal{K}_2} \hat{q}_k^n(t)$ for some strict nonempty subsets $\mathcal{K}_1 \subset \mathcal{K}$ and $\mathcal{K}_2 = \mathcal{K} \setminus \mathcal{K}_1$ (that is, the set $\mathcal{K}_1$ consists of queues that are strictly shorter than those not in it at time t), then

$$\sum_{k\in\mathcal{K}_1} \frac{d}{dt} \bar{\phi}_k^n(\lambda^n(t - \frac{\hat{u}^n(0)}{n})) \geq \left(\frac{\mu_{\mathcal{K}_1}}{\mu_{\mathcal{K}}} + \sigma\right) \lambda^n, \quad \text{for } t \geq \frac{1}{n}, \tag{183}$$

where σ is chosen as any positive number satisfying $\sigma \leq [1-(\mu_{\mathcal{K}_2}/\mu_{\mathcal{K}})^c] - \mu_{\mathcal{K}_1}/\mu_{\mathcal{K}}$ (for BR policy with c servers randomly selected upon each job arrival), and the function $\bar{\phi}_k^n(\lambda^n(t - \hat{u}^n(0)/n))$, an item in (175), is the amount of fluid flowing into the server k by time t. As there are finite number of non-empty strict subsets of $\mathcal{K}$, the constant σ can be chosen independent of the sets $\mathcal{K}_1$ and $\mathcal{K}_2$.

Consider any (regular) time $t > 1/n$ and $\hat{q}_{\max}^n(t) > 0$ (i.e., $|\hat{q}^n(t)| > 0$). There are two possible cases. First, if $\mathcal{K}_{\max}^t$ is a strict subset of $\mathcal{K}$, then by letting $\mathcal{K}_2 = \mathcal{K}_{\max}^t$, we have from the property in (183),

$$\sum_{k\in\mathcal{K}_{\max}^t} \frac{d}{dt} \bar{\phi}_k^n(\lambda^n(t - \frac{\hat{u}^n(0)}{n})) \leq \sum_{k\in\mathcal{K}_{\max}^t} \frac{\mu_k}{\mu_{\mathcal{K}}} \lambda^n.$$

Observe that $\hat{q}_k^n(t) > 0$ for $k \in \mathcal{K}_{\max}^t$, which implies

$$\dot{\bar{b}}_k^n(t) = 1, \quad \text{for } k \in \mathcal{K}_{\max}^t. \tag{184}$$

Therefore, summing up (175) over $k \in \mathcal{K}_{\max}^t$ and taking derivative yield,

$$\begin{aligned} \sum_{k\in\mathcal{K}_{\max}^t} \dot{\hat{q}}_k^n(t) &= \sum_{k\in\mathcal{K}_{\max}^t} n\left(\frac{d}{dt}\bar{\phi}_k^n(\lambda^n(t - \frac{\hat{u}^n(0)}{n})) - \mu_k\right) \leq \sum_{k\in\mathcal{K}_{\max}^t} n\left(\frac{\mu_k}{\mu_{\mathcal{K}}}\lambda^n - \mu_k\right) \\ &= \sum_{k\in\mathcal{K}_{\max}^t} \frac{\mu_k}{\mu_{\mathcal{K}}} n(\lambda^n - \mu_{\mathcal{K}}) = \sum_{k\in\mathcal{K}_{\max}^t} \frac{\mu_k}{\mu_{\mathcal{K}}}(\theta_{\mathcal{K}} + o(1)) \leq -\kappa_3, \end{aligned} \tag{185}$$

for some constant $\kappa_3 > 0$ that is independent of the index n (which is sufficiently large) and the set $\mathcal{K}_{\max}^t$. (Here, $o(1)$ stands for a quantity that converges to 0 as $n \to \infty$.) Given the property in (182), the above implies

$$\dot{\hat{q}}_{\max}^n(t) = \frac{1}{K_{\max}^t} \sum_{k\in\mathcal{K}_{\max}^t} \dot{\hat{q}}_k^n(t) \leq -\frac{\kappa_3}{K}.$$

In the other case, if $\mathcal{K}_{\max}^t = \mathcal{K}$ (and $K_{\max}^t = K$), then by summing up (175) for all $k \in \mathcal{K}$ and taking derivative, we have

$$\hat{q}_{\mathcal{K}}^n(t) = \hat{q}_{\mathcal{K}}^n(0) + n(\lambda^n(t - \frac{\hat{u}^n(0)}{n})) - n\sum_{k\in\mathcal{K}} \mu_k \left(\bar{b}_k^n(t) - \frac{\hat{v}_k^n(0)}{n}\right),$$

and then

$$\dot{\hat{q}}^n_{\mathcal{K}}(t) = n\lambda^n - n\sum_{k\in\mathcal{K}}\mu_k = \theta_{\mathcal{K}} + o(1) \le -\kappa_4,$$

for some constant $\kappa_4 > 0$ that is independent of the (sufficiently large) index n. Consequently, we have

$$\dot{\hat{q}}^n_{\max}(t) = \frac{1}{K}\dot{\hat{q}}^n_{\mathcal{K}}(t) \le -\frac{\kappa_4}{K}.$$

In summary, we can find a constant $\kappa_5 > 0$ such that whenever $\hat{q}^n_{\max}(t) > 0$, we have $\dot{\hat{q}}^n_{\max}(t) \le -\kappa_5$ for $t > 1/n$, for sufficiently large n. This implies the conclusion in (173) given the bound in (177). □

Proof of Theorem 7 (outline). The stability of the fluid model corresponding to the n-th system $\hat{\Xi}^n(t)$, just established in Lemma 19(b), directly implies the positive recurrence of the Markov process $\hat{\Xi}^n(t)$, as well as the existence and uniqueness of its stationary distribution, according to Theorem 4.2 of Dai [17]. The finiteness of the $(p-1)$-th moment of the queue length and the convergence in (120) follow from Theorem 4.1(ii) of Dai and Meyn [19]. □

A.6 Preparation for Proof of Theorems 8 and 9

Before proceeding, we introduce two variations of the diffusion limit theorems that will play an auxiliary role in the proofs of Theorems 8 and 9. The first one allows more flexible initial states in the sequence of systems $\{\hat{\Xi}^n(t)\}$.

Proposition 20 Suppose the sequence of initial states $\{\hat{\Xi}^n(0); n \in \mathcal{N}\}$ is tight. Let $\{t^n_0; n \in \mathcal{N}\}$ be any sequence of times such that $t^n_0 \to 0$ and $nt^n_0 \to \infty$ as $n \to \infty$. Then, for any subsequence of $\mathcal{N}$, there exists a further subsequence, denoted by $\mathcal{N}_1$, such that the following weak convergence holds when $n \to \infty$ along $\mathcal{N}_1$:

$$\hat{\Xi}^k(t^n_0 + t) = \left(\hat{Q}^n(t^n_0 + t), \hat{U}^n(t^n_0 + t), \hat{V}^n(t^n_0 + t), \hat{\Psi}^n(t^n_0 + t)\right) \Rightarrow \left(\hat{Q}(t), 0, 0, 0\right),$$

where the limit $\hat{Q}(t)$ follows the specifications in Theorem 3(a) (resp. Theorem 5(a), Theorem 6(a), Proposition 2(a)) under the $RR(\theta)$ policy (resp. $AC(\theta, h)$, $SC(\theta, h)$, JSQ/BR policy). Furthermore, we have for any $M \ge 0$,

$$\limsup_{n\to\infty, n\in\mathcal{N}_1} \mathsf{P}\{\kappa|\hat{\Xi}^n(0)| \le M\} \quad \le \quad \mathsf{P}\{|\hat{Q}(0)| \le M\}, \tag{186}$$

where κ is a constant that depends only on network parameters.

To describe the second variation of the diffusion limit theorem, we introduce a linear Skorohod problem which is obtained by removing the randomness in the diffusion limit given in (32-34): for each $k \in \mathcal{K}$,

$$\hat{q}_k(t) = \hat{q}_k(0) + \theta_k t + \hat{y}_k(t) \ge 0, \tag{187}$$

$$\int_0^\infty \hat{q}_k(s)d\hat{y}_k(s)ds = 0, \tag{188}$$

$$\hat{y}_k(t) \text{ is non-decreasing with } \hat{y}^n_k(0) = 0. \tag{189}$$

The processes $\hat{q}_k(t)$ and $\hat{y}_k(t)$ satisfying the above constitute a one-dimensional reflecting mapping (Lemma 10). Clearly, as $\theta_k < 0$ (for any admissible policy), it is stable:

$$\hat{q}_k(t) = 0, \quad \text{for } t \ge \frac{\hat{q}_k(0)}{-\theta_k}. \tag{190}$$

In the second variation, we add an extra scaling to the diffusion-scaled systems, and relate the sequence of systems with "mixed" scalings to the linear Skorohod problem just described.

Proposition 21 Let $\{m_n; n \in \mathcal{N}\}$ be any sequence that increases to infinity (i.e., $m_n \to \infty$ as $n \to \infty$), and assume that the sequence of initial states $\{\hat{\Xi}^n(0)/m_n, n \in \mathcal{N}\}$ is tight. Let $\{t_0^n; n \in \mathcal{N}\}$ be any sequence of times such that $t_0^n \to 0$ and $nt_0^n \to \infty$ as $n \to \infty$. Then, for any subsequence of $\mathcal{N}$, there exists a further subsequence, denoted by $\mathcal{N}_1$, such that the following weak convergence holds when $n \to \infty$ along $\mathcal{N}_1$:

$$\begin{aligned}
&\frac{1}{m_n}(\hat{X}^n(m_n(t_0^n+t)) - \hat{X}^n(m_n t_0^n)) \Rightarrow \theta t,\\
&\frac{1}{m_n}\left(\hat{U}^k(m_n(t_0^n+t)), \hat{V}^k(m_n(t_0^n+t)), \hat{\Psi}^k(m_n(t_0^n+t))\right) \Rightarrow 0,\\
&\frac{1}{m_n}\left(\hat{Q}^k(m_n(t_0^n+t)), \hat{Y}^n(m_n(t_0^n+t)) - \hat{Y}^n(m_n t_0^n)\right) \Rightarrow (\hat{q}(t), \hat{y}(t)),
\end{aligned} \tag{191}$$

where the limit $(\hat{q}(t), \hat{y}(t))$ follows the specifications in (187-189) under the $RR(\theta)$, $AC(\theta, h)$, $SC(\theta, h)$, or JSQ/BR (with $\theta_k = \theta_{\mathcal{K}}/K$) policy. Furthermore, we have for any $M \geq 0$,

$$\limsup_{n\to\infty, n\in\mathcal{N}_1} \mathsf{P}\{\kappa|\hat{\Xi}^n(0)/m_n| \leq M\} \quad \leq \quad \mathsf{P}\{|\hat{q}(0)| \leq M\}, \tag{192}$$

where κ is a constant that depends only on network parameters.

Proposition 20 can be proved following the idea outlined by Bramson (page 115 of [9], also see [43, 52]). In Ye and Yao [66], we also extend the diffusion limit theorem, for a more complex resource-sharing network, to allow for a tight initial sequence of Markovian states as in Proposition 20 here, and provide a detailed proof following Bramson's idea. That proof can be adapted to prove the proposition here, and hence we omit the details. The proof of Proposition 21 is a straightforward modification of the arguments that establish the diffusion limit in Proposition 20 (cf. the proof of Proposition 3(b) of [66]), and hence is also omitted.

The stability properties described in (173,190) (for the limits in Lemma 19 and Proposition 21 respectively) can be turned into a kind of pathwise stability of the respective systems, which will lead to a stronger moment stability.

Lemma 22 There exists a time t_0 such that the following conclusions hold.
(a) Let $\{m_i; i = 1, 2, \cdots\}$ be a sequence of number such that $m_i \to \infty$ as $i \to \infty$; and let $\{x^i \in \mathcal{X}; i = 1, 2, \cdots\}$ be a sequence of initial states such that $|x^i| \leq m_i$ for all i. Then, for any sufficiently large $n \in \mathcal{N}$, the following holds (with probability one), as $i \to \infty$,

$$\frac{1}{m_i}\left(\hat{Q}^n(m_i t; x^i), \hat{\Psi}^n(m_i t; x^i)\right) \to 0 \quad \text{ u.o.c. of } t \geq t_0.$$

(b) Let $\{m_n; n \in \mathcal{N}\}$ be a sequence of numbers such $m_n \to \infty$ as $n \to \infty$; and assume that the sequence of initial states $\{\hat{\Xi}^n(0)\}$ satisfies $|\hat{\Xi}^n(0)| \leq m_n$. Then, the following holds (with probability one): as $n \to \infty$ (along the full sequence $\mathcal{N}$),

$$\frac{1}{m_n}\left(\hat{Q}^n(m_n t), \hat{\Psi}^n(m_n t)\right) \to 0 \quad \text{ u.o.c. of } t \geq t_0.$$

Two conclusions in the above lemma can be established by applying the stability properties (173) and (190) to the limits in (171) and (191), respectively. The proof repeats the one for Lemma 10(a,b) of [66] and is omitted.

A.7 Proof of Theorem 8: Interchange of Limits under RR, AC, SC and JSQ Policies

In the next two lemmas, we establish a pathwise bound and a moment bound for the Markov states of the sequence of systems. The moment bound (bounded p-th moment) will play two

bridging roles. First, it ensures the uniform integrability required for converting the pathwise stability property in Lemma 22 to some moment stability property in Lemma 26 below. Second, it restricts the variability of the system when we follow the approach in Budhiraja and Lee [11] (a refinement of the one in Dai and Meyn [19]) to bound a return time — a key step to establish a required tightness property leading the interchange of limits. (The proof related to this second role will be omitted as it almost repeats the corresponding one for Proposition 12 of [66]. Refer to that proof for details.)

Lemma 23 (Pathwise bound) Suppose $RR(\theta)$, $AC(\theta, h)$, $SC(\theta, h)$ or JSQ policy is in force. There exists a constant $\kappa > 0$ such that for any index n and any time $t \geq 0$, the following bounds hold,

$$|\hat{\Psi}^n(t)| \leq \kappa \left(1 + |\hat{\Psi}^n(0)| + \sup_{0\leq s\leq t} |\hat{E}^n(s)|\right), \tag{193}$$

$$|\hat{Q}^n(t)| \leq \kappa \left(1 + |\hat{\Xi}^n(0)| + \sup_{0\leq s\leq t} |\hat{E}^n(s)| + \sum_{k\in\mathcal{K}} \sup_{0\leq s\leq t} |\hat{S}^n_k(s)| + t\right). \tag{194}$$

Proof. *For RR, AC and SC policies.* Consider $AC(\theta, h)$ policy first. Fix an index n and a time $t > 0$ arbitrarily. We first establish the following bound concerning the routing process: for some constant κ_1 (independent of t and k),

$$|\Phi^n_k(E^n(t)) - p^n_k E^n(t))| \leq \kappa_1 \left(1 + |\Psi^n(0)| + \sup_{0\leq s\leq t} |E^n(s) - \lambda^n s|\right). \tag{195}$$

Pick any server k. Let τ be the maximum time, before time t, such that the surplus of routing deviation is less than 0,

$$\tau = \sup\{s : \Psi^n_k(s) \leq 0,\ 0 \leq s \leq t\};$$

if the set at the right-hand-side is empty, we let $\tau = 0$. Consider the case with $\tau < t$ (and ignore the trivial case with $\tau = t$). Observe that for any time $s \in (\tau, t]$, we have $\Psi^n_k(s) > 0$. Since the sum $\sum_{k'\in\mathcal{K}} \Psi^n_{k'}(s) = 0$ (from (113)), the smallest value of $\Psi^n_{k'}(s)$, $k' \in \mathcal{K}$, must be negative and be attained with some k' other than class-k. Hence, according to the AC policy defined in Section 6.1, no job will be assigned to class-k queue during the time interval $(\tau, t]$, which implies

$$\Phi^n_k(E^n(t')) = \Phi^n_k(E^n(\tau)), \quad \text{for } t' \in (\tau, t].$$

Then, we have for any $t' \in (\tau, t]$,

$$\begin{aligned} & \Phi^n_k(E^n(t')) - p^n_k E^n(t')) \leq \Phi^n_k(E^n(\tau)) - p^n_k E^n(\tau)) \\ = \ & \Psi^n_k(\tau) - \Psi^n_k(0) + h^n_k(E^n(\tau) - \lambda^n \tau), \end{aligned}$$

where the equality follows from (111). Note that from the definition of τ, we have $\Psi^n_k(\tau-) \leq 0$ and thus $\Psi^n_k(\tau) \leq 1$ if $\tau > 0$. (Recall the assumption of continuous interarrival times. Without loss of generality, we can assume that any job's arrival does not coincide with any other arrival or service completion event. Thus, there is at most one arrival at time τ.) Therefore, if $\tau > 0$, the above gives

$$\Phi^n_k(E^n(t')) - p^n_k E^n(t')) \leq 1 + |\Psi^n_k(0)| + \sup_{0\leq s\leq t'} |h^n_k(E^n(s) - \lambda^n s)|, \quad \text{for } t' \in (\tau, t]. \tag{196}$$

Clearly, the above estimate still holds if $\tau = 0$. Now, as the above estimate is valid for any k and particularly for $t' = t$, we have

$$\begin{aligned} & -\Phi^n_k(E^n(t)) - p^n_k E^n(t)) = \sum_{k'\neq k, k'\in\mathcal{K}} [\Phi^n_{k'}(E^n(t)) - p^n_{k'} E^n(t))] \\ \leq \ & (K-1) + |\Psi^n(0)| + \sum_{k'\in\mathcal{K}} \sup_{0\leq s\leq t'} |h^n_{k'}(E^n(s) - \lambda^n s)|. \end{aligned} \tag{197}$$

The bounds in (196,197) implies (195) with properly chosen constant κ_1.

Now, from (111,195), we have

$$|\Psi^n(t)| \le \kappa_2 \left(1 + |\Psi^n(0)| + \sup_{0\le s\le t} |E^n(s) - \lambda^n s|\right). \tag{198}$$

As $(\hat{Q}_k^n(t), Y_k^n(t))$ constitutes a one-dimensional Skorohod mapping (refer to (20,21,9,10) or (22-24)), we have (cf. Lemma 10 and inequality (128)),

$$\begin{aligned} Q_k^n(t) &= Q_k^n(0) + X_k^n(t) + \sup_{0\le s\le t} [-Q_k^n(0) - X_k^n(s)]^+ \\ &\le 2Q_k^n(0) + 2 \sup_{0\le s\le t} |X_k^n(s)|. \end{aligned}$$

Using the expression in (21) and the estimate in (195) in the above yield the following bound,

$$Q_k^n(t) \le \kappa_3 \left(1 + Q_k^n(0) + |\Psi^n(0)| + \sup_{0\le s\le t} |E^n(s) - \lambda^n s| + \sup_{0\le s\le t} |S_k^n(s) - \mu_k s| + \frac{1}{n}t\right). \tag{199}$$

Finally, by applying diffusion scaling to the two bounds in (198,199) and choosing the constant κ properly, we establish the bounds in (193,194) under $AC(\theta, h)$ policy.

The proof for $SC(\theta, h)$ simply repeats the above, with the item $h_k^n(E^n(t) - \lambda^n t)$ in the chasing process of the AC policy being replaced by

$$\delta_k^n(t) := (S_k^n(B_k^n(t)) - \mu_k B_k^n(t)) - \sum_{i\in\mathcal{K}} h_{ki}^n (S_i^n(B_i^n(t)) - \mu_i B_i^n(t)),$$

and keeping in mind that $B_k^n(t'') - B_k^n(t') \le t'' - t'$ for all $0 \le t' \le t''$. The proof for the $RR(\theta)$ policy is immediate since it is special case of the $AC(\theta, h)$ or $SC(\theta, h)$ policy.

For JSQ policy. Fixed an index n and a time $t > 0$ arbitrarily, and consider any (fixed) sample path. Observe that the queue length process $\hat{Q}^n(s)$ is a piecewise constant RCLL function of time $s \ge 0$. Denote the time of successive job arrivals as $t_i = \Upsilon^n(i)$ for convenience. As the interarrival times follow a continuous distribution with mean $1/\lambda^n > 0$, we can assume without loss of generality: (1) $0 = t_0 \le t_1 < t_2 < t_3 < \cdots$ and $t_i \to \infty$ (recall, $t_1 = u^n(1)$ is the initial residual arrival time), (2) the time t_i ($i \ge 2$) does not coincide with any service completion time as well. Hence, at time t_i ($i \ge 2$), there is an arrival, and if it is routed to, say, server k, then this arrival triggers an increase in queue k, $\hat{Q}_k^n(t_i) = \hat{Q}_k^n(t_i-) + 1/n$, while no change in other queues.

Denote the (diffusion scaled) maximum queue length process and its maximum during time interval $[0, t]$ as

$$\hat{Q}_{\max}^n(s) := \max_{k\in\mathcal{K}} \hat{Q}_k^n(s), \quad q^* := \max_{s\in[0,t]} \hat{Q}_{\max}^n(s). \tag{200}$$

Observe that the maximum value q^* must be attained at a job arrival time (and for a time period thereafter). Let t_m be the first time at which the maximum queue length attains the value q^* within time interval $[0, t]$:

$$m = \min\{i : \hat{Q}_{\max}^n(t_i) = q^*,\ t_i \le t\}.$$

If $t_m = t_0$ or t_1, the bound in (194) is satisfied immediately. Hence, without loss of generality, assume $t_m > 0$ (or $m \ge 2$). Let k^* be the index of longest queue at time t_m, and we have

$$\hat{Q}_{k^*}^n(t_m) = \hat{Q}_{k^*}^n(t_m-) + \frac{1}{n} = q^*. \tag{201}$$

Observe that all queue lengths are at most $(q^* - 1/n)$ just before time t_m, and that class-k^* queue length is among the shortest ones just before t_m. Therefore, we have

$$\hat{Q}_k(t_m) = \hat{Q}_k(t_m-) = q^* - \frac{1}{n}, \quad k \neq k^*. \tag{202}$$

Similar to (200), denote the minimum queue length as

$$\hat{Q}^n_{\min}(s) := \min_{k\in\mathcal{K}} \hat{Q}^n_k(s).$$

Let t_ℓ be the minimum time such that no queue is empty during $[t_\ell, t_m]$:

$$\ell = \min\{i : \hat{Q}^n_{\min}(s) > 0 \text{ for } s \in [t_i, t_m]\}.$$

Then, all servers are busy during $[t_\ell, t_m]$: for all $k \in \mathcal{K}$,

$$\hat{Y}^n_k(s) = \hat{Y}^n_k(t_\ell), \quad s \in [t_\ell, t_m]. \tag{203}$$

Now, from the property in (203) and the dynamics in (22,25), we have

$$\begin{aligned}\sum_{k\in\mathcal{K}} \hat{Q}^n_k(t_m) &= \sum_{k\in\mathcal{K}} \hat{Q}^n_k(t_\ell) + (\hat{E}^n(t_m) - \hat{E}^n(t_\ell)) \\ &\quad - \sum_{k\in\mathcal{K}} (\hat{S}^n_k(\tilde{B}^n_k(t_m)) - \hat{S}^n_k(\tilde{B}^n_k(t_\ell))) + \theta^n_{\mathcal{K}}(t_m - t_\ell).\end{aligned} \tag{204}$$

From (201,202), the left-hand-side of the above is,

$$\sum_{k\in\mathcal{K}} \hat{Q}^n_k(t_m) = q^* + (K-1)(q^* - \frac{1}{n}). \tag{205}$$

If $t_\ell = t_0$, then the equality in (204) is reduced to

$$Kq^* - \frac{K-1}{n} = |\hat{Q}^n_k(0)| + \hat{E}^n(t_m) - \sum_{k\in\mathcal{K}} \hat{S}^n_k(\tilde{B}^n_k(t_m)) + \theta^n_{\mathcal{K}} t_m.$$

Keeping in mind that $|\hat{Q}^n(t)| \leq Kq^*$, the above implies (194). On the other hand, if $t_\ell > t_0$, from the definition of t_ℓ, we must have

$$\hat{Q}^n_{\min}(t_\ell-) = 0 \text{ and } \hat{Q}^n_{\min}(t_\ell) = \frac{1}{n}, \tag{206}$$

which implies that a job must arrive at time t_ℓ and join the unique empty queue. Hence, at time t_ℓ, at least one of the queues is equal to $1/n$ while all other queues are at most q^* (under diffusion scaling). Consequently, we have

$$\sum_{k\in\mathcal{K}} \hat{Q}^n_k(t_\ell) \leq \frac{1}{n} + (K-1)q^*. \tag{207}$$

Putting (207,205) into (204) yields

$$q^* \leq \frac{K}{n} + (\hat{E}^n(t_m) - \hat{E}^n(t_\ell)) - \sum_{k\in\mathcal{K}} (\hat{S}^n_k(\tilde{B}^n_k(t_m)) - \hat{S}^n_k(\tilde{B}^n_k(t_\ell))) + \theta^n_{\mathcal{K}}(t_m - t_\ell), \tag{208}$$

which implies (194) as well.

Recall $\hat{\Psi}^n(t; JSQ) \equiv 0$. Thus, the bound about the chasing process in (193) is trivial and can be omitted for the JSQ policy. □

Under the p-moment bound of the primitives in (117), the pathwise bound just established is converted to a moment bound of the system state in the next lemma. The proof is same as the one for Lemma 9(a) of [66]; the additional routing component, $\hat{\Psi}^n(t)$, makes nearly no difference to the proof given the bound in (193). The detailed proof is omitted.

Lemma 24 (Bounded p-th moment) Suppose $RR(\theta)$, $AC(\theta,h)$, $SC(\theta,h)$ or JSQ policy is in force, and the p-th moment condition in (117) holds. Then, the following holds for some constant κ,

$$\mathsf{E}\sup_{0\le s\le t}|\hat{Q}^k(s)|^p+\mathsf{E}\sup_{0\le s\le t}|\hat{\Psi}^n(s)|^p\le\kappa(|\hat{\Xi}^n(0)|^p+1+t^p); \tag{209}$$

and consequently (redefining κ), for any $0\le q\le p$,

$$\mathsf{E}\sup_{0\le s\le t}|\hat{Q}^k(s)|^q+\mathsf{E}\sup_{0\le s\le t}|\hat{\Psi}^n(s)|^q\le\kappa(|\hat{\Xi}^n(0)|^q+1+t^q).$$

With the bounded p-th moment of the system state in the above lemma, the pathwise stability properties in Lemma 22 can be turned into corresponding moment stability properties in the next lemma. The latter serves as intermediate steps to establish the uniform p-th moment stability (Proposition 26) that holds the key to establishing the tightness of the stationary distributions associated with the sequence of systems $\{\hat{\Xi}^n(t), n\in\mathcal{N}\}$.

Lemma 25 Suppose $RR(\theta)$, $AC(\theta,h)$, $SC(\theta,h)$ or JSQ policy is in force, and the p-th moment condition in (117) holds. Then, there exists a time t_0 such that the following conclusions hold. (a) Assume $\{m_i\}$ and $\{x^i\}$ as in Lemma 22(a). Then, the followings hold for sufficiently large n,

$$\{|\frac{1}{m_i}\hat{Q}^n(m_it;x^i)|^p\}\text{ and }\{|\frac{1}{m_i}\hat{\Psi}^n(m_it;x^i)|^p\}\text{ are uniformly integrable (w.r.t. } i),\text{ for any } t\ge 0,$$

and

$$\lim_{i\to\infty}\mathsf{E}\frac{1}{m_i^p}\left(\left|\hat{Q}^n(m_it;x^i)\right|^p+\left|\hat{\Psi}^n(m_it;x^i)\right|^p\right)=0,\quad\text{for any } t\ge t_0. \tag{210}$$

(b) Assume $\{m_n\}$ and $\{\hat{\Xi}^n(0)\}$ as in Lemma 22(b). Then, the followings hold,

$$\{|\frac{1}{m_n}\hat{Q}^n(m_nt)|^p\}\text{ and }\{|\frac{1}{m_n}\hat{\Psi}^n(m_nt)|^p\}\text{ is uniformly integrable (w.r.t. } n),\quad\text{for any } t\ge 0,$$

and

$$\lim_{n\to\infty}\mathsf{E}\frac{1}{m_n^p}\left(\left|\hat{Q}^n(m_nt)\right|^p+\left|\hat{\Psi}^n(m_nt)\right|^p\right)=0,\quad\text{for } t\ge t_0. \tag{211}$$

The proof of the uniform integrability properties in the above lemma is same as the one for Lemma 9(b,c) of [66] and is omitted. Then, these uniform integrability properties justify the interchange of the expectation and the limit in (210) and (211) respectively, and therefore the conclusions in (210) and (211) follow from Lemma 22 immediately.

Next, by putting the results in (210) and (211) together, we can show the uniform p-th moment stability property whose proof is same as the one for Proposition 11 of [66] and hence omitted.

Proposition 26 (Uniform p-th moment stability) Suppose $RR(\theta)$, $AC(\theta,h)$, $SC(\theta,h)$ or JSQ policy is in force, and the p-th moment condition in (117) holds. Then, there exists a time t_0 and a sufficiently large index n_0 such that the following holds for all $t\ge t_0$,

$$\lim_{|x|\to\infty}\sup_{n\ge n_0}\mathsf{E}\frac{1}{|x|^p}\left(\left|\hat{Q}^n(|x|t;x)\right|^p+\left|\hat{\Psi}^n(|x|t;x)\right|^p\right)=0. \tag{212}$$

Now, we are able to establish the tightness and a uniform p-th moment bound of the sequence of steady states $\{\hat{\Xi}^n(\infty), n \in \mathcal{N}\}$, by making use of moment bound and moment stability results in Properties 24 and 26 and following the approach developed in [11, 19]. Refer to [66] as well, and the detailed proof is omitted. Finally, given such tightness and moment bound properties of the steady states, the proof of our interchange of limits in Theorem 8 (for RR, AC, SC and JSQ policies) is identical to the proof of Theorem 4 in [66] (Proposition 20, a variation of diffusion limit theorem, is used here), which is a modification of the standard argument leading to the interchange of limits from the tightness in [11, 24, 25, 36].

A.8 Proof of Theorem 9: Interchange-of-Limits under BR Policy

As a preparation, we first introduce a new fluid model and its uniform attraction property. This fluid model is associated with the whole sequence of systems $\{\hat{\Xi}^n(t), n \in \mathcal{N}\}$, which is derived as the limit (or cluster point) of a sequence of systems under the hydrodynamic scaling (cf. the uniform continuity in Lemma 29). Intuitively, it can also be obtained from a critically loaded system (i.e., $\rho = 1$ in the system described in (5-10) and under BR routing), by removing the randomness of arrival and service processes while taking into account the initial residuals. It is described by the following set of conditions:

$$\begin{aligned}
\bar{q}_k(t) &= \bar{q}_k(0) + \bar{\phi}_k\left(\lambda(t - t \wedge \bar{u}(0))\right) - \mu_k\left(\bar{b}_k(t) - \bar{b}_k(t) \wedge \bar{v}_k(0)\right) \\
&= \bar{q}_k(0) + \bar{\phi}_k\left(\lambda(t - t \wedge \bar{u}(0))\right) - \mu_k\left(t - \bar{b}_k(t) \wedge \bar{v}_k(0)\right) + \bar{y}_k(t) \geq 0, \qquad (213) \\
&\bar{\phi}_k(t) \text{ is non-decreasing with } \bar{\phi}_k(0) = 0, \text{ and Lipschitz with constant 1,} \qquad (214) \\
&\bar{b}_k(t) \text{ is non-decreasing with } \bar{b}_k(0) = 0, \text{ and Lipschitz with constant 1,} \qquad (215) \\
&\bar{y}^n_k(t) = \mu_k(t - \bar{b}_k(t)) \text{ is non-decreasing with } \bar{y}_k(0) = 0, \qquad (216) \\
&\int_0^\infty \bar{q}_k(s) d\bar{y}_k(s) ds = 0. \qquad (217)
\end{aligned}$$

In addition, if $\max_{k\in\mathcal{K}_1} \bar{q}_k(t) < \min_{k\in\mathcal{K}_2} \bar{q}_k(t)$ for some strict nonempty subset $\mathcal{K}_1 \subset \mathcal{K}$ and $\mathcal{K}_2 = \mathcal{K} \setminus \mathcal{K}_1$ (that is, the set $\mathcal{K}_1$ consists of queues that are strictly shorter than those not in it at time t), then

$$\sum_{k\in\mathcal{K}_1} \frac{d}{dt}\bar{\phi}_k(\lambda(t - t \wedge \bar{u}(0))) \geq \left(\frac{\mu_{\mathcal{K}_1}}{\mu_{\mathcal{K}}} + \sigma_\phi\right)\lambda, \quad \text{for } t \geq \bar{u}(0), \qquad (218)$$

where σ_ϕ is any positive number independent of $(\mathcal{K}_1, \mathcal{K}_2)$ and satisfying $\sigma_\phi \leq [1 - (\mu_{\mathcal{K}_2}/\mu_{\mathcal{K}})^c] - \mu_{\mathcal{K}_1}/\mu_{\mathcal{K}}$.

The above inequality indicates that the shorter queues in the (critically loaded) fluid model described above will receive flows more than the capacity of associated servers, which also implies that the longer queues get less. Consequently, all short and long queues converge to the same length. This observation leads to the uniform attraction property in the following proposition, which is a key to establishing a bound for the queue length (in Lemma 30).

Proposition 27 (Theorem 4 of [16]) (a) There exists a constant $\kappa_w > 0$ that only depends on the network parameters such that for any solution to the fluid model in (213-218) satisfying $|\bar{q}(0)| + |\bar{u}(0)| + |\bar{v}(0)| \leq 1$, the following bound holds,

$$|\bar{q}(t)| \leq \kappa_w, \quad t \geq 0. \qquad (219)$$

(b) (Uniform attraction) There exists a time T_0 such that the following holds for any solution to the fluid model in (213-218) satisfying $|\bar{q}(0)| + |\bar{u}(0)| + |\bar{v}(0)| \leq 1$:

$$\bar{q}_1(t) = \cdots = \bar{q}_K(t) = q^*, \quad \text{for } t \geq T_0, \qquad (220)$$

where q^* is a constant.

The above proposition (part (b) in particular) is a variation of Theorem 4 of [16]. The new feature here, the additional residuals $\bar{u}(0)$ and $\bar{v}(0)$, can be dealt with in the same way as in the proof of Lemma 19(b) so that the effect of these residuals can be mitigated after the time $t = 1$. Hence, the proof is omitted.

Next, we will introduce the regular events and derive a probabilistic bound for these events.

With the initial residuals u_1^n and $v_{k,1}^n$ removed from $E^n(t)$ and $S_k^n(t)$, the (undelayed) arrival and service processes are denoted: $E^{o,n}(t)$ and $S^{o,n}(t) = (S_k^{o,n}(t))_{k\in\mathcal{K}}$, $t \geq 0$, where

$$E^{o,n}(t) = \max\left\{i : \sum_{\ell=2}^{i} u_\ell^n \leq t\right\}, \quad \text{and} \quad S_k^{o,n}(t) = \max\left\{i : \sum_{\ell=2}^{i} v_{k,\ell}^n \leq t\right\}. \tag{221}$$

Here and below, the superscript "o" denotes the undelayed version of a (possibly) delayed renewal process.

Define the variables:

$$u^{n,\max}(t) := \max\left\{u_i^n : \sum_{\ell=2}^{i-1} u_\ell^n \leq t,\ i = 2, 3, \cdots\right\}, \tag{222}$$

$$v_k^{n,\max}(t) := \max\left\{v_{k,i}^n : \sum_{\ell=2}^{i-1} v_{k,\ell}^n \leq t,\ i = 2, 3, \cdots\right\}. \tag{223}$$

The first variable is the maximal interarrival time of jobs realized before time t for the n-th system; the second variable is analogous, for the service times. Note that the initial residuals u_1^n and $v_{k,1}^n$ are excluded.

Let t^* and u^* be any positive times, and $\{m_n, n \in \mathcal{N}\}$ be a sequence of real numbers with $m_n \geq 1$. Define the regular events as:

$$\Omega^n(t^*, u^*, m_n) = \Omega_u^n(t^*, m_n) \cap \Omega_v^n(t^*, m_n) \cap \Omega_E^n(t^*, u^*, m_n) \cap \Omega_S^n(t^*, u^*, m_n) \cap \Omega_X^n(t^*, m_n) \tag{224}$$

where

$$\Omega_u^n(t^*, m_n) = \left\{\frac{1}{nm_n} u^{n,\max}(n^2 m_n t^*) \leq \frac{1}{n^{(p^*-2)/2p^*}}\right\}, \tag{225}$$

$$\Omega_v^n(t^*, m_n) = \bigcap_{k\in\mathcal{K}} \left\{\frac{1}{nm_n} v_k^{n,\max}(n^2 m_n t^*) \leq \frac{1}{n^{(p^*-2)/2p^*}}\right\}, \tag{226}$$

$$\Omega_E^n(t^*, u^*, m_n) = \left\{\sup_{0\leq t\leq nt^*} \sup_{0\leq u\leq u^*} |\frac{1}{m_n}(\bar{E}^{o,n}(m_n(t+u)) - \bar{E}^{o,n}(m_n t)) - \lambda^n u| \leq \frac{1}{\log n}\right\}, \tag{227}$$

$$\Omega_S^n(t^*, u^*, m_n) = \left\{\sup_{0\leq t\leq nt^*} \sup_{0\leq u\leq u^*} |\frac{1}{m_n}(\bar{S}^{o,n}(m_n(t+u)) - \bar{S}^{o,n}(m_n t)) - \mu u| \leq \frac{1}{\log n}\right\}, \tag{228}$$

$$\Omega_X^n(t^*, m_n) = \left\{\sup_{0\leq t\leq t^*} \frac{1}{m_n}(|\hat{E}^{o,n}(m_n t)| + |\hat{S}^{o,n}(m_n t)|) \leq \frac{n}{\log n}\right\}. \tag{229}$$

The ranges that bound the sample paths are carefully specified such that the probabilities of these events must approach one at a certain rate as indicated in the following lemma (Lemma 3.1 of [67]).

Lemma 28 Suppose the p^*-th moment condition in (121) holds, and let t^* and u^* be any positive times. Then, the following estimate holds for sufficiently large n (depending on t^* and u^*),

$$\mathsf{P}(\Omega^n(t^*, u^*, m_n)) \geq 1 - \frac{(\log n)^{p^*+1}}{n^{p^*/2-2}}, \quad \text{for all } m_n \geq 1.$$

We remark that in the above regular events and the hydrodynamic scaling to be defined immediately, the parameters $\{m_n, n \in \mathcal{N}\}$ are included so that the system under a "mixed" scaling, $\hat{\Xi}^n(m_n t)/m_n$, can be analyzed directly. These parameters also serve to contain the sequence of initial states $\{\hat{\Xi}^n(0)\}$ when we allow $|\hat{\Xi}^n(0)|$ to increase infinitely in the hydrodynamic analysis below.

Now, we introduce the hydrodynamics representation for our model, and establish a uniform continuity property (Lemma 29) in which the fluid model described in (213-218) is derived as the limit point for the hydrodynamic processes restricted to the regular events.

The hydrodynamics here is a modification of the original one by Bramson [9], and is dedicated to studying the system under a "mixed" scaling, $\hat{\Xi}^n(m_n t)/m_n$. Denote

$$y^n[= y^n(\omega, \Delta, m_n)] := \max\left(\frac{1}{m_n}|\hat{Q}^n(0)| + \sup_{0\le t\le\Delta}\frac{1}{m_n}|\hat{X}^n(m_n t)|,\ \frac{1}{m_n}|\hat{\Xi}^n(0)|,\ 1\right), \quad (230)$$

for any time interval $[0, \Delta]$ (for the process $\hat{\Xi}^n(m_n t)/m_n$), with $\Delta > 0$, and any sequence of numbers $\{m_n \ge 1; n \in \mathcal{N}\}$. Let $T > 0$ be a fixed time of a certain magnitude (to be specified later). Divide the time interval $[0, m_n\Delta]$ (for the process $\hat{\Xi}^n(t)$) into a total of $\lceil n\Delta/y^nT\rceil$ segments with equal length y^nm_nT/n, where $\lceil\cdot\rceil$ denotes the integer ceiling. The j-th segment, $j = 0, ..., \lceil n\Delta/y^nT\rceil - 1$, covers the time interval $[jy^nm_nT/n, (j+1)y^nm_nT/n]$ (of $\hat{\Xi}^n(t)$). Note that the last interval (with $j = \lceil n\Delta/y^nT\rceil - 1$) covers a negligible piece of time beyond the right end of $[0, \Delta]$ if $n\Delta/y^nT$ is not an integer. For simplicity, below we shall treat $n\Delta/y^nT$ as an integer so as to omit the ceiling notation. Then, for any $t \in [0, \Delta]$, we can write $t = y^nm_n(jT + u)/n$ for some $j = 0, \cdots, n\Delta/y^nT$ and $u \in [0, T]$. Therefore, for $u \in [0, T]$ and $j \le n\Delta/y^nT$, we write

$$\begin{aligned}\frac{1}{y^nm_n}\hat{Q}^n(m_nt) &= \frac{1}{y^nm_n}\hat{Q}^n(\frac{jy^nm_nT + y^nm_nu}{n})\\ &= \frac{1}{ny^nm_n}Q^n(ny^nm_n(jT+u)) := \bar{Q}^{n,j}(u). \end{aligned}\quad (231)$$

Hence, this hydrodynamic scaling can be regarded as a special fluid scaling with the scaling parameter ny^nm_n and a shifted start time. The processes, $\bar{\Xi}^{n,j}(u)$, $\bar{U}^{n,j}(u)$ and $\bar{V}^{n,j}(u)$, are defined in the same manner. The arrival, routing, busy time and service processes are written as,

$$\bar{E}^{n,j}(u) := \frac{1}{ny^nm_n}[E^n(ny^nm_n(jT+u)) - E^n(ny^nm_njT)], \quad (232)$$

$$\bar{\Phi}^{n,j}_k(u) := \frac{1}{ny^nm_n}[\Phi^n_k(ny^nm_n(\bar{\bar{E}}^{n,j} + u)) - \Phi^n_k(ny^nm_n\bar{\bar{E}}^{n,j})], \quad (233)$$

$$\bar{B}^{n,j}_k(u) := \frac{1}{ny^nm_n}[B^n_k(ny^nm_n(jT+u)) - B^n_k(ny^nm_njT)], \quad (234)$$

$$\bar{S}^{n,j}_k(u) := \frac{1}{ny^nm_n}[S^n_k(ny^nm_n(\bar{\bar{B}}^{n,j}_k + u)) - S^n_k(ny^nm_n\bar{\bar{S}}^{n,j}_k)], \quad (235)$$

where

$$(\bar{\bar{E}}^{n,j}, \bar{\bar{S}}^{n,j}_k) := \frac{1}{ny^nm_n}\big(E^n(ny^nm_njT), B^n_k(ny^nm_njT)\big). \quad (236)$$

It can be verified that for the routing and service components in the system equation in (25), we have

$$\bar{\Phi}^{n,j}_k(\bar{E}^{n,j}(u)) := \frac{1}{y^nm_n}\Big(\hat{\Phi}^n_k\big(\tilde{E}^n(\frac{y^nm_n}{n}(jT+u))\big) - \hat{\Phi}^n_k\big(\tilde{E}^n(\frac{y^nm_n}{n}jT)\big)\Big), \quad (237)$$

$$\bar{S}^{n,j}_k(\bar{B}^{n,j}_k(u)) := \frac{1}{y^nm_n}\Big(\hat{S}^n_k\big(\tilde{B}^n_k(\frac{y^nm_n}{n}(jT+u))\big) - \hat{S}^n_k\big(\tilde{B}^n_k(\frac{y^nm_n}{n}jT)\big)\Big). \quad (238)$$

Lemma 29 (Uniform continuity) Let M, Δ (and $\Delta^* := \Delta + 1$), T and T^* be any given positive numbers, and suppose $|\hat{\Xi}^n(0)| \vee 1 \leq m_n$ for all n. (Δ and T are the parameters in the definition of the hydrodynamic processes above, and Δ^* and T^* will be used to specify the regular event.) (a) For any $\epsilon > 0$, there exists n^* such that for any $n \geq n^*$, the following holds for any $\omega \in \Omega^n(\Delta^*, T^*, m_n)$ and $0 \leq j \leq n\Delta/y^n T$: if

$$|\bar{Q}^{n,j}(0)| + |\bar{U}^{n,j}(0)| + |\bar{V}^{n,j}(0)| \leq M, \tag{239}$$

then, we can find a fluid model $(\bar{q}(t), \bar{u}(0), \bar{v}(0))$ satisfying (213-218) and $|\bar{q}(0)| + |\bar{u}(0)| + |\bar{v}(0)| \leq M$ such that

$$\sup_{0 \leq u \leq T} |\bar{Q}^{n,j}(u) - \bar{q}(u)| + |\bar{U}^{n,j}(0) - \bar{u}(0)| + |\bar{V}^{n,j}(0) - \bar{v}(0)| < \epsilon. \tag{240}$$

(b) Moreover, the time T can be chosen sufficiently long (depending on network parameters only) such that the following holds for any $\omega \in \Omega^n(\Delta^*, T^*, m_n)$ and $1 \leq j \leq n\Delta/y^n T$ (excluding $j = 0$):

$$\bar{U}^{n,j}(0) \text{ and } \bar{V}^{n,j}(0) \leq \frac{1}{k^{(p^*-1)/2p^*}}.$$

Consequently, for any $\epsilon > 0$, there exists n^* such that the following holds for any $n \geq n^*$, $|\hat{\Xi}^n(0)| \vee 1 \leq m_n$, $\omega \in \Omega^n(\Delta^*, T^*, m_n)$, and $1 \leq j \leq n\Delta/y^n T$, if

$$|\bar{Q}^{n,j}(0)| \leq M, \tag{241}$$

then, we can find a fluid model $(\bar{q}(t), \bar{u}(0) = 0, \bar{v}(0) = 0)$ satisfying (213-218) and $|\bar{q}(0)| \leq M$ such that

$$\sup_{0 \leq u \leq T} |\bar{Q}^{n,j}(u) - \bar{q}(u)| < \epsilon.$$

In [67], a similar result for multiclass queueing networks is established. In that paper, a tedious process is used to decompose the system equations into the corresponding fluid model plus the associated random components, and to show the random components vanish uniformly. Then, a general uniform continuity property (also established there) is invoked to conclude the required result. In contrast, we will prove the above lemma directly, through a contradictory argument involving the conventional fluid limit. This approach is more transferable to similar studies.

Proof. Suppose the conclusion does not hold. That is, there exists an $\epsilon_0 > 0$ and a subsequence of index $\mathcal{N}_1 \subset \mathcal{N}$ such that for any $n \in \mathcal{N}_1$, we can find a sample $\omega^n \in \Omega^n(\Delta^*, T^*, m_n)$ and an integer $j_n \in [0, n\Delta/y^n T]$ for which the condition in (239) holds but the bound in (240) does not: that is,

$$|\bar{Q}^{n,j_n}(0)| + |\bar{U}^{n,j_n}(0)| + |\bar{V}^{n,j_n}(0)| \leq M, \tag{242}$$

and, for any fluid model $(\bar{q}(t), \bar{u}(0), \bar{v}(0))$ satisfying (213-218) and $|\bar{q}(0)| + |\bar{u}(0)| + |\bar{v}(0)| \leq M$,

$$\sup_{0 \leq u \leq T} |\bar{Q}^{n,j_n}(u) - \bar{q}(u)| + |\bar{U}^{n,j_n}(0) - \bar{u}(0)| + |\bar{V}^{n,j_n}(0) - \bar{v}(0)| \geq \epsilon_0. \tag{243}$$

(Note that the processes $\bar{Q}^{n,j_n}(u)$, $\bar{U}^{n,j_n}(u)$ and $\bar{V}^{n,j_n}(u)$ specified just now are sample paths associated with ω_n.)

Applying the conventional approach for proving fluid limit theorem (refer to for example [13, 17], among many others), however, we can find a further subsequence $\mathcal{N}_2 \subset \mathcal{N}_1$ such that as $n \to \infty$ along $\mathcal{N}_2$, we have

$$\sup_{0 \leq u \leq T} |\bar{Q}^{n,j_n}(u) - \bar{q}(u)| + |\bar{U}^{n,j_n}(0) - \bar{u}(0)| + |\bar{V}^{n,j_n}(0) - \bar{v}(0)| \to 0, \tag{244}$$

for some fluid model $(\bar{q}(t), \bar{u}(0), \bar{v}(0))$ satisfying (213-218) with $|q(0)|+|u(0)|+|v(0)| \leq M$, which contradicts to (243).

It is left to claim the limit $(\bar{q}(t), \bar{u}(0), \bar{v}(0))$ in the above convergence. Using the hydrodynamic scaling described in (231-238), the system equations in (6-10) (indexed by n properly) can be rewritten as,

$$\bar{Q}_k^{n,j}(t) = \bar{Q}_k^{n,j}(0) + \bar{\Phi}_k^{n,j}(\bar{E}^{n,j}(t)) - \bar{S}_k^{n,j}(\bar{B}_k^{n,j}(t)) \geq 0, \tag{245}$$

$$\bar{B}_k^{n,j}(t) = \int_0^t 1_{\{\bar{Q}_k^{n,j}(s)>0\}} ds, \tag{246}$$

$$\bar{Y}_k^{n,j}(t) = \mu_k(t - \bar{B}_k^{n,j}(t)) = \mu_k \int_0^t 1_{\{\bar{Q}_k^{n,j}(s)=0\}} ds, \tag{247}$$

$$\int_0^\infty \bar{Q}_k^{n,j}(s) d\bar{Y}_k^{n,j}(s) = 0, \tag{248}$$

$$\bar{Y}_k^{n,j}(t) \text{ is non-decreasing in } t \geq 0, \text{ and } \bar{Y}_k^{n,j}(0) = 0. \tag{249}$$

From the condition (242), we can find a subsequence $\mathcal{N}_2 \subset \mathcal{N}_1$ such that

$$(\bar{Q}^{n,j_n}(0), \bar{U}^{n,j_n}(0), \bar{V}^{n,j_n}(0)) \to (\bar{q}(0), \bar{u}(0), \bar{v}(0)), \tag{250}$$

with $|\bar{q}(0)| + |\bar{u}(0)| + |\bar{v}(0)| \leq M$.

Observe that $\bar{E}^{n,j}(s)$ is a (scaled) delayed renewal process, and its undelayed version $\bar{E}^{o,n,j}(s)$ can be defined in a way similar to (221). It can be verified directly the two versions satisfy the following,

$$\bar{E}^{n,j}(s) = \frac{1_{\{s \geq \bar{U}^{n,j}(0)\}}}{ny^n m_n} + \bar{E}^{o,n,j}(s - s \wedge \bar{U}^{n,j}(0)). \tag{251}$$

Recall from the proof of Lemma 3.4 of [67] (the bound in (B.16) in particular) that under the condition $|\hat{\Xi}^n(0)| \leq m_n$ and $\omega_n \in \Omega^n(\Delta^*, T^*, m_n)$, the following convergence holds as $n \to \infty$ (along the full sequence $\mathcal{N}$),

$$\sup_{0 \leq j < \frac{n\Delta}{y^n T}} \sup_{s \in [0,T]} |\bar{E}^{o,n,j}(s - s \wedge \bar{U}^{n,j}(0)) - \lambda^n(s - s \wedge \bar{U}^{n,j}(0))| \to 0.$$

(As a side note, the initial condition, $|\hat{\Xi}^n(0)| \leq m_n$, imposes certain bound on the parameter y^n used in defining the hydrodynamic processes. Such a bound is required when we employ the property of the regular event to claim the above convergence. Refer to [67] for technical details.) Given the relationship in (251), the above convergence gives the following immediately,

$$\sup_{s \in [0,T]} |\bar{E}^{n,j_n}(s) - \lambda(s - s \wedge \bar{u}(0))| \to 0, \quad \text{as } n \to \infty \text{ along } \mathcal{N}_2. \tag{252}$$

Similarly, we have

$$\sup_{s \in [0,T]} |\bar{S}_k^{n,j_n}(s) - \mu_k(s - s \wedge \bar{v}_k(0))| \to 0, \quad \text{as } n \to \infty \text{ along } \mathcal{N}_2. \tag{253}$$

Given (250,252,253), we are now able to apply the conventional approach (as in the proof of Lemma 19) to derive the fluid limit for $\bar{Q}_k^{n,j}(t)$ as depicted in (244). □

The uniform attraction and the uniform continuity (in Lemmas 27 and 29) established above are concerned with properties of the system sequence in $O(n)$-long period under the original scaling (here, ignore the additional scalling factors y^n and m_n for simplicity). Assembling these properties across $O(n^2)$-long period (or, constant period under the diffusion scaling), we can bound the queue length over regular events in the following lemma.

Lemma 30 Consider any time interval $[0, \Delta]$, with $\Delta > 0$, and suppose $|\hat{\Xi}^n(0)| \vee 1 \leq m_n$ for all n. Let $\epsilon > 0$ be any given (small) number. Then, there exists a sufficiently large T such that for sufficiently large n, the following results hold for any $\omega \in \Omega^n(\Delta^*, T^*, m_n)$ (here $\Delta^* = \Delta + 1$ and $T^* = 2T$) and *positive* integers $j = 1, \cdots, n\Delta/y^n T$:
(a) (Uniform attraction)

$$|\bar{Q}_k^{n,j}(u) - \frac{1}{K}\bar{Q}_{\mathcal{K}}^{n,j}(u)| \leq \epsilon, \quad \text{for all } u \in [0,T],\ k \in \mathcal{K}. \tag{254}$$

(b) (Complementarity) If $\bar{Q}_{\mathcal{K}}^{n,j}(u') > 2K\epsilon$ for some $u' \in [0,T]$, then

$$\bar{Y}_{\mathcal{K}}^{n,j}(u) - \bar{Y}_{\mathcal{K}}^{n,j}(0) = 0, \quad \text{for all } u \in [0,T]. \tag{255}$$

(c) (Boundedness)

$$|\bar{Q}^{n,j}(u)| \leq \kappa, \quad \text{for all } u \in [0,T], \tag{256}$$

where κ is a positive constant that depends only on system parameters (independent of n and ω). In addition, the bound in (256) also applies to $j = 0$.

The above lemma is indeed a version of Lemma 3.5 of [67], but in the context of the parallel server model here. The proof in that paper is transferable with minor modifications (e.g., carefully handling certain complementarity property). As the model under study is simpler in the network structure, the proof can be substantially simplified, which is included as follows.

Proof. Let T be any real value satisfying:

$$T \quad \geq \quad \kappa T_0, \tag{257}$$

where the time T_0 is given in Lemma 27(b). Note that T is large enough so that in the fluid model in Theorem 27 (under the heavy traffic condition), the state $\bar{q}(t)$ will reach the fixed-point state (satisfying $\bar{q}_1(t) = \cdots = \bar{q}_K(t)$), starting from an initial state $(\bar{q}(0), \bar{u}(0), \bar{v}(0))$ that is bounded by κ. Here κ is a constant that depends on network parameters only:

$$\kappa = 2\kappa_w + 5, \tag{258}$$

where κ_w is given in Lemma 27. The rationale for the choice of κ will become evident shortly.

We prove the lemma in two steps.

Step 1. Prove (a,b) for $j = 1$ and (c) for $j = 0, 1$.

Note that $|\bar{\Xi}^{n,0}(0)| = |\hat{\Xi}^n(0)/y^n m_n| \leq 1$ according to the definitions in (230,231). By Lemma 29, we have for sufficiently large n, and $\omega \in \Omega^n(\Delta^*, T^*, m_n)$, there exists a fluid model $(\bar{q}(u), \bar{u}(0), \bar{v}(0))$ satisfying (213-218) which may depend on n, $\hat{\Xi}^n(0)$, m_n and ω, such that,

$$|\bar{q}(0)| + |\bar{u}(0)| + |\bar{v}(0)| \leq 1, \quad \text{and} \tag{259}$$

$$\sup_{u \in [0,T^*]} \left(|\bar{Q}^{n,0}(u) - \bar{q}(u)| + |\bar{U}^{n,0}(0) - \bar{u}(0)| + |\bar{V}^{n,0}(0) - \bar{v}(0)| \right) < \frac{\epsilon}{2}. \tag{260}$$

Since $T \geq \kappa T_0$ $(\geq T_0)$, applying the uniform attraction property in Lemma 27 to the above fluid model $\bar{q}(u)$ yields:

$$\bar{q}_1(u) = \cdots = \bar{q}_K(u) = q^*, \quad \text{for all } u \geq T, \quad \text{and} \tag{261}$$

$$|\bar{q}(u)| \leq \kappa_w, \quad \text{for all } u \geq 0. \tag{262}$$

Note that $\bar{Q}^{n,0}(T + u) \equiv \bar{Q}^{n,1}(u)$. Then, it follows from (260,261) that the conclusion (a) holds with $j = 1$ for sufficiently large n and for all $\omega \in \Omega^n(\Delta^*, T^*, m_n)$.

By (259,262) again, we have for all $u \in [0, T^*]$ $(= [0, 2T])$,

$$|\bar{Q}^{n,0}(u)| \le |\bar{q}(u)| + \frac{\epsilon}{2} \le \kappa_w + \frac{\epsilon}{2},$$

and for all $u \in [0, T]$,

$$|\bar{Q}^{n,1}(u)| = |\bar{Q}^{n,0}(T + u)| \le \kappa_w + \frac{\epsilon}{2} \ (\le \kappa). \tag{263}$$

That is, the bounding property in (c), for both $j = 0$ and $j = 1$, is satisfied.

When $\bar{Q}^{n,1}_{\mathcal{K}}(u') > 2K\epsilon$ for some $u' \in [0, T]$, we have from (254) (for $j = 1$),

$$\bar{Q}^{n,1}_k(u') \ge \frac{\bar{Q}^{n,1}_{\mathcal{K}}(u')}{K} - \epsilon > \epsilon,$$

which, along with (260,261), implies

$$q^* > \bar{Q}^{n,1}_k(u') - \frac{\epsilon}{2} \ge \frac{\epsilon}{2},$$

and thus

$$\bar{Q}^{n,1}_k(u) > q^* - \frac{\epsilon}{2} > 0, \quad \text{for all } u \in [0, T], \ k \in \mathcal{K}.$$

Therefore, following the complementarity relationship in (248), we have,

$$\bar{Y}^{n,1}_k(u) \text{ does not increase in } u \in [0, T], \text{ for all } k \in \mathcal{K}.$$

In summary, if $\bar{Q}^{n,1}_{\mathcal{K}}(u') > K\epsilon$ for some $u' \in [0, T]$, then,

$$\bar{Y}^{n,1}_{\mathcal{K}}(u) \text{ does not increase in } u \in [0, T];$$

that is, the conclusion (b) holds for $j = 1$.

Step 2. We now extend the above to $j = 2, \ldots, n\Delta/y^n T$.

Suppose again to the contrary, there exists a subsequence $\mathcal{N}_1 \subset \mathcal{N}$ such that, for any $n \in \mathcal{N}_1$, at least one of the results in (a,b,c) does not hold for some integer $j \in [2, n\Delta/y^n T]$ and some sample-path $\omega \in \Omega^n(\Delta^*, T^*, m_n)$. Let j_n be the smallest positive integer in the interval $[2, n\Delta/y^n T]$ such that at least one of the properties in (a,b,c) does not hold with the associated $\hat{\Xi}^n(0)$, m_n and ω. To reach a contradiction, in the rest of the proof we will show that the desired properties in (a,b,c) hold for $j = j_n$ for sufficiently large $n \in \mathcal{N}_1$, and indeed for any $\omega \in \Omega^n(\Delta^*, T^*, m_n)$.

Following the earlier argument, under the (contradictory) assumption above, the results in (a,b,c) hold for $j = 1, ..., j_n - 1$ and any $\omega \in \Omega^n(\Delta^*, T^*, m_n)$, for each $n \in \mathcal{N}_1$. Specifically, for $j = j_n - 1$ (≥ 1), we have

$$|\bar{Q}^{n,j_n-1}(0)| \le \kappa, \quad \text{for all } n \in \mathcal{N}_1.$$

By Lemma 29(b), we have for any sufficiently large $n \in \mathcal{N}_1$ and any $\omega \in \Omega^n(\Delta^*, T^*, m_n)$, there exists a fluid model $(\bar{q}(u), \bar{u}(0) = 0, \bar{v}(0) = 0)$ satisfying (213-218) (which may depend on n, $\hat{\Xi}^n(0)$, m_n and ω) such that

$$\sup_{u \in [0, T^*]} |\bar{Q}^{n,j_n-1}(u) - \bar{q}(u)| < \frac{\epsilon}{2}, \tag{264}$$

with $|\bar{q}(0)| \le \kappa$. (Here, we know that $|\bar{U}^{n,j_n-1}(0)| + |\bar{V}^{n,j_n-1}(0)| \to 0$ as $n \to 0$, and can set $\bar{u}(0) = \bar{v}(0) = 0$ by Lemma 29(b).) Since $T \ge \kappa T_0$, applying the uniform attraction property in Lemma 27 to the above limit yields the following again:

$$\bar{q}_1(u) = \cdots = \bar{q}_K(u) = q^*, \quad \text{for all } u \ge T. \tag{265}$$

Note that $\bar{Q}^{n,j_n-1}(T+u) \equiv \bar{Q}^{n,j_n}(u)$. Hence, the bound in (264), along with (265), implies that the conclusion (a) holds with $j = j_n$ and $\omega \in \Omega^n(\bar{\Delta}^*, T^*, m_n)$, for sufficiently large $n \in \mathcal{N}_1$.

Following the same procedure for proving the conclusion (b) for $j = 1$, we can show that the complementarity property in conclusion (b) holds for $j = j_n$.

Now, consider any sufficiently large $n \in \mathcal{N}_1$, such that the results in (a,b) hold for $j = 1, \cdots, j_n$ (but it needs not holds for $j = 0$) and for all $\omega \in \Omega^n(\Delta^*, T^*, m_n)$. This implies that when restricted to $\omega \in \Omega^n(\Delta^*, T^*, m_n)$, the processes,

$$(q(t), x(t), y(t)) := \frac{1}{y^n m_n}(\hat{Q}^n_{\mathcal{K}}(m_n t), \hat{Q}^n_{\mathcal{K}}(\frac{m_n y^n T}{n}) + \hat{X}^n_{\mathcal{K}}(m_n t) - \hat{X}^n_{\mathcal{K}}(\frac{m_n y^n T}{n}), \hat{Y}^n_{\mathcal{K}}(m_n t)),$$

satisfy the specifications in Lemma 12(b) *in the time interval* $[y^n T/n, (j_n y^n T + y^n T)/n]$, which merges all intervals $[jy^n T/n, (j+1)y^n T/n]$ for $j = 1, \cdots, j_n$. Hence, we have for any $t \in [y^n T/n, (j_n y^n T + y^n T)/k]$ $(\subset [0, \Delta^*])$,

$$\begin{aligned}
& \frac{1}{y^n m_n}\hat{Q}^n_{\mathcal{K}}(m_n t) - 2K\epsilon \\
\leq\ & \Phi\left(\frac{1}{y^n m_n}\left(\hat{Q}^n_{\mathcal{K}}(\frac{m_n y^n T}{n}) + \hat{X}^n_{\mathcal{K}}(m_n \cdot) - \hat{X}^n_{\mathcal{K}}(\frac{m_n y^n T}{n})\right) - 2K\epsilon\right)(t) \\
\leq\ & 2 \sup_{s \in [\frac{y^n T}{n}, t]} \left| \frac{1}{y^n m_n}\left(\hat{Q}^n_{\mathcal{K}}(\frac{m_n y^n T}{n}) + \hat{X}^n_{\mathcal{K}}(m_n s) - \hat{X}^n_{\mathcal{K}}(\frac{m_n y^n T}{n})\right) - 2K\epsilon \right| \\
\leq\ & 2(\kappa_w + \frac{\epsilon}{2} + 2 + 2K\epsilon),
\end{aligned}$$

where in the second inequality we have also applied the conclusion in (263) (i.e., $\hat{Q}^n_{\mathcal{K}}(m_n y^n T/n)/y^n m_n = |\bar{Q}^{n,1}(0)| \leq \kappa_w + \epsilon/2$) and the definition in (230). Keeping in mind that $\bar{Q}^{n,j_n}(u) \equiv \hat{Q}^n((j_n y^n m_n T + y^n m_n u)/n)/y^n m_n$, the above implies that (c) holds with $j = j_n$ for sufficiently large $n \in \mathcal{N}_1$. □

The rest of the proof for Theorem 9 follows the same procedure in [67], starting from Lemma 3.6 of that paper. It is also the procedure in the proof of Theorem 8 (for RR, AC, SC and JSQ policies), from Lemma 24 onward in Section A.5. We omit the detailed proof and outline the procedure as follows.

First, we employ the pathwise bound of the queue length process for sample path in the regular event (Lemma 30(c)) and the probabilistic bound of the regular event (Lemma 28), to establish the *bounded p-th moment of queue length*:

$$\mathsf{E} \sup_{0 \leq s \leq t} \left| \frac{1}{m_n}\hat{Q}^n(m_n s) \right|^p \quad \leq \quad \kappa(1 + t^p), \tag{266}$$

for sufficiently large n and for some constant $\kappa > 0$. Note, in Lemma 24, we have established a similar inequality in (209) for the RR, AC, SC and JSQ policies without requiring the stronger p^*-th moment condition in (121), since the queue length process (and the associated chasing process as well) under these policies can be bounded by the "free" primitive processes (Lemma 23).

Second, using the p-th moment bound of queue length in (266), we turn the pathwise stability results in Lemma 22 to a *uniform p-th moment stability*:

$$\lim_{|x| \to \infty} \sup_{n \geq n_0} \mathsf{E} \frac{1}{|x|^p} \left| \hat{Q}^n(|x|t; x) \right|^p = 0, \quad t \geq t_0, \tag{267}$$

for some time t_0 and some (large) index n_0. This is parallel to Proposition 26.

Finally, given the above moment bound and stability of the queue length process, we are able to establish the tightness of the stationary distributions of the pre-limit queue length processes and thus the interchange of limits, the main result in Theorem 9. Refer to the comments following Proposition 26 (for RR, AC, SC and JSQ policies) as well.

References

[1] Altman, E., B. Gaujal and A. Hordijk. (2003). *Discrete-Event Control of Stochastic Networks: Multimodularity and Regularity*, Springer, Heidelberg.

[2] Atar, R., A. Mandelbaum and M.I. Reiman. (2004). Scheduling a multiclass queue with many exponential servers: Asymptotic optimality in heavy traffic. *Annals of Applied Probability*, **14**(3), 1084-1134.

[3] Badonnel, R. and M. Burgess. (2008). Dynamic pull-based load balancing for autonomic servers. *IEEE Network Operations and Management Symposium*, 751-754.

[4] Banawan, S. A. and Zeidat, N. M. (1992). A comparative study of load sharing in heterogeneous multicomputer systems. In *Proceedings of 25th Annual Simulation Symposium* (pp. 22-31). IEEE.

[5] Bell, S.L., and R.J. Williams. (2001). Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: asymptotic optimality of a threshold policy. *Annals of Applied Probability*, **11**(3), 608-649.

[6] der Boor, M. V., Borst, S. C., Van Leeuwaarden, J. S., Mukherjee, D. (2022). Scalable load balancing in networked systems: A survey of recent advances. *SIAM Review*, **64**(3), 554-622.

[7] Borst, S., A. Mandelbaum and M.I. Reiman. (2004). Dimensioning large call centers. *Operations Research*, **52**(1), 17-34.

[8] Bramson, M. (1998). Stability of two families of queueing networks and a Discussion of Fluid Limits. *Queueing Systems: Theory and Applications*, **23**, 7-31.

[9] Bramson, M. (1998). State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems, Theory and Applications*. **30**, 89-148.

[10] Bramson, M. (2011). Stability of join the shortest queue networks. *Annals of Applied Probability*, **21**(4), 1568-1625.

[11] Budhiraja, A. and C. Lee. (2009). Stationary distribution convergence for generalized Jackson networks in heavy traffic. *Mathematics of Operations Research*, **34**(1), 45-56.

[12] Chao, X., L. Liu and S. Zheng. (2003). Resource allocation in multisite service systems with intersite customer flows. *Management Science*, **49**(12), 1739-1752.

[13] Chen, H. (1995). Fluid approximations and stability of multiclass queueing networks: work-conserving discipline. *Annals of Applied Probability*, **5**, 637-655.

[14] Chen, H. and J.G. Shanthikumar. (1994). Fluid limits and diffusion approximations for networks of multi-server queues in heavy traffic. *Journal of Discrete Event Dynamic Systems*, **4**, 269-291.

[15] Chen, H. and D.D. Yao. (2001). *Fundamentals of Queueing Networks: Performance, Asymptotics and Optimization*, Springer, New York.

[16] Chen, H. and H.Q. Ye. (2012). Asymptotic optimality of balanced routing. *Operations Research*, **60**(1), 163-179.

[17] Dai, J.G. (1995). On positive Harris recurrence of multi-class queueing networks: a unified approach via fluid limit models. *Annals of Applied Probability*, **5**, 49-77.

[18] Dai, J.G. and W. Lin. (2008). Asymptotic optimality of maximum pressure policies in stochastic processing networks, *Annals of Applied Probability*, **18**, 2239-2299.

[19] Dai, J.G. and S.P. Meyn. (1995). Stability and convergence of moments for multiclass queueing networks via fluid models. *IEEE Transactions on Automatic Control*, **40**, 1899-1904.

[20] Dai, J.G. and G. Weiss. (1996). Stability and instability of fluid models for reentrant lines. *Mathematics of Operations Research*, **21**, 115-134.

[21] Davis, M.H.A. (1984). Piecewise-deterministic Markov processes: a general class of nondiffusion models. *Journal of the Royal Statistical Society, Series B*, **46**, 353-388.

[22] Eschenfeldt, P. and D. Gamarnik. (2018). Join the shortest queue with many servers: the heavy-traffic asymptotics. *Mathematics of Operations Research*, **43**(3), 867-886.

[23] Foss, S. and Stolyar, A. L. (2017). Large-scale join-idle-queue system with general service times. *Journal of Applied Probability*, **54**(4), 995-1007.

[24] Gamarnik, D. and A. Zeevi. (2006). Validity of heavy traffic steady-state approximations in Generalized Jackson Networks. *Annals of Applied Probability*, **16**, 56-96.

[25] Gurvich, I. (2014). Validity of heavy-traffic steady-state approximations in multiclass queueing networks: the case of queue-ratio disciplines. *Mathematics of Operations Research*, **39**, 121-162.

[26] Gupta, V., M.H. Balter, K. Sigman and W. Whitt. (2007). Analysis of join-the-shortest-queue routing for web server farms. *Performance Evaluation*, **64**(9-12), 1062-1081.

[27] Gut, A. (2005). *Probability: A Graduate Course*, Springer, New York.

[28] Hajek, B. (1985). External splittings of point processes. *Mathematics of Operations Research*, **10**(4), 543-556.

[29] Halfin, S. and W. Whitt. (1981) Heavy-traffic limits for queues with many exponential servers. *Operations Research* **29**(3):567–588.

[30] Harrison, J.M. (1996). The BIGSTEP approach to flow management in stochastic processing networks. In *Stochastic Networks: Theory and Applications* (F.P. Kelly, S. Zachary and I. Ziedins, eds.), 57-90. Oxford Univ. Press.

[31] Harrison, J.M. (1998). Heavy traffic analysis of a system with parallel servers: asymptotic analysis of discrete-review policies. *Annals of Applied Probability*, **8**, 822-848.

[32] Harrison, J.M. and M. J. López. (1999). Heavy traffic resource pooling in parallel-server systems. *Queueing Systems: Theory and Applications*, **33**, 339-368.

[33] Harrison, J.M. and Wein, L.M. (1990). Scheduling networks of queues: Heavy traffic analysis of a two-station closed network. *Operations research*, **38**(6), 1052-1064.

[34] Hurtado-Lange, D. and S.T. Maguluri. (2020). Transform methods for heavy-traffic analysis. *Stochastic Systems*, **10**(4), 275-309.

[35] Hyytia, E. and S. Aalto. (2016). On Round-Robin routing with FCFS and LCFS scheduling. *Performance Evaluation*, **97**, 83-103.

[36] Katsuda, T. (2010). State-space collapse in stationarity and its application to a multiclass single-server queue in heavy traffic. *Queueing Systems: Theory and Applications*, **65**, 237-273.

[37] Kelly, F.P. and C.N. Laws. (1993). Dynamic routing in open queueing networks: Brownian models, cut constraints and resource pooling. *Queueing Systems: Theory and Applications*, **13**, 47-86.

[38] Kobayashi, M., Y. Sakuma and M. Miyazawa. (2013). Join the shortest queue among k parallel queues: tail asymptotics of its stationary distribution. *Queueing Systems: Theory and Applications*, **74**(2-3), 303-332.

[39] Lee, C., A.R. Ward and H.Q. Ye. (2020). Stationary distribution convergence of the offered waiting processes for GI/GI/1+GI queues in heavy traffic. *Queueing Systems: Theory and Applications*, **94**(1), 147-173.

[40] Lee, C., A.R. Ward and H.Q. Ye. (2021). Stationary distribution convergence of the offered waiting processes in heavy traffic under general patience time scaling. *Queueing Systems: Theory and Applications*, **99**(3-4), 283-303..

[41] Liu, Z. and R. Righter. (1998). Optimal load balancing on distributed homogeneous unreliable processors. *Operations Research*, **46**(4), 563-573.

[42] Liu, Z. and D. Towsley. (1994). Optimality of the Round Robin routing policy. *Journal of Applied Probability*, **31**(2), 466-475.

[43] Mandelbaum, A. and A.L. Stolyar. (2004). Scheduling flexible servers with convex delay costs: heavy-traffic optimality of the generalized $c\mu$-rule. *Operations Research*, **52**, 836-855.

[44] Martins, L.F., S.E. Shreve and H.M. Soner. (1996). Heavy traffic convergence of a controlled, multiclass queueing system. *SIAM journal on control and optimization*, **34**(6), 2133-2171.

[45] Mitzenmacher, M. (2001). The power of two choices in randomized load balancing. *IEEE Transactions on Parallel and Distributed Computing*, **12**, 1094-1104.

[46] Moyal, P. and Perry, O. (2022). Stability of parallel server systems. *Operations Research*, **70**(4), 2456-2476.

[47] Reiman, M.I. (1984). Some diffusion approximations with state space collapse. *Modelling and Performance Evaluation Methodology*, F. Baccelli and G. Fayolle (editors), Springer-Verlag, 209-240.

[48] Rybko, A.N. and A.L. Stolyar. (1992). Ergodicity of stochastic processed describing the operations of open queueing networks. *Problemy Peredachi Informatsii*, **28**, 2-26.

[49] Selen, J., Adan, I., Kapodistria, S., van Leeuwaarden, J. (2016). Steady-state analysis of shortest expected delay routing. *Queueing Systems*, **84**(3), 309-354.

[50] Shiryaev, A.N. (1996). *Probability* (2ed), Springer-Verlag, New York.

[51] Stolyar, A.L. (1994). On the stability of multiclass queueing network. *Proceeding of the 2nd International Conference on Telecommunication System–Moleling and Analysis*, Nashville, Tennessee, 23-35.

[52] Stolyar, A.L. (2004). Max-weight scheduling in a generalized switch: state space collapse and workload minimization in heavy traffic. *Annals of Applied Probability,* **14**, 1-53.

[53] Stolyar, A. L. (2005). Optimal routing in output-queued flexible server systems. *Probability in the Engineering and Informational Sciences*, **19**(2), 141-189.

[54] Stolyar, A.L. (2015). Pull-based load distribution in large-scale heterogeneous service systems. *Queueing Systems: Theory and Applications*, **80**(4), 341-361.

[55] Tsoukatos, K.P. and A.M. Makowski. (2006). Asymptotic optimality of the Round-Robin policy in multipath routing with resequencing. *Queueing Systems: Theory and Applications*, **52**, 199-214.

[56] Weber, R.R. (1978). On the optimal assignment of customers to parallel servers. *Journal of Applied Probability*, **15**(2), 406-413.

[57] Whitt, W. (1986). Deciding which queue to join: some counterexamples. *Operations Research*, **34**(1), 55-62.

[58] Whitt, W. (2002). *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*, Springer, New York.

[59] Williams, R.J. (1998b). Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing Systems, Theory and Applications*, **30**(1-2), 27-88.

[60] Williams, R.J. (2000). On dynamic scheduling of a parallel server system with complete resource pooling. *Analysis of Communication Networks: Call Centres, Traffic and Performance*, D.R. McDonald and S.R.E. Turner (eds.), Fields Institute Communications Volume 28, American Mathematical Society, pp.49-71.

[61] Winston, W. (1977). Optimality of the shortest line discipline. *Journal of Applied Probability*, **14**(1), 181-189.

[62] Whitt, W. (1992). Understanding the efficiency of multi-server service systems. *Management Science*. **38**, 708-723.

[63] Wu, R. and D.G. Down. (2009). Round robin scheduling of heterogeneous parallel servers in heavy traffic. *European Journal of Operational Research*, **195**, 372-380.

[64] Ye, H. and D.D. Yao. (2008). Heavy traffic optimality of a stochastic network under utility-maximizing resource control. *Operations Research*, **56**(2), 453-470.

[65] Ye, H.Q. and D.D. Yao. (2012). A stochastic network under fair resource control — diffusion limit with multiple bottlenecks. *Operations Research*, **60**(3), 716-738.

[66] Ye, H.Q. and D.D. Yao. (2016). Diffusion limit of fair resource control — stationary and interchange of limits. *Mathematics of Operations Research*, **41**(4), 1161-1207.

[67] Ye, H.Q. and D.D. Yao. (2018). Justifying diffusion approximations for multiclass queueing networks under a moment condition. *Annals of Applied Probability*, **28**(6), 3652-3697.

[68] Zhang, H.Q., G.H. Hsu and R. Wang. (1995). Heavy traffic limit theorems for sequence of shortest queueing systems. *Queueing Systems: Theory and Applications*, **21**, 217-238.