Queueing Syst (2018) 88:27–71
https://doi.org/10.1007/s11134-017-9549-7

Many-server scaling of the N-system under FCFS–ALIS

Dongyuan Zhan[1] · **Gideon Weiss**[2]

Received: 10 March 2016 / Revised: 19 September 2017 / Published online: 9 October 2017

Abstract The N-system with independent Poisson arrivals and exponential server-dependent service times under the first come first served and assign to the longest idle server policy has an explicit steady-state distribution. We scale the arrival rate and the number of servers simultaneously, and obtain the fluid and central limit approximation for the steady state. This is the first step toward exploring the many-server scaling limit behavior of general parallel service systems.

Keywords N-system · Many-server scaling · Fluid limits · Central limits · First come first served · Assign to the longest idle server

Mathematics Subject Classification 60K25

1 Introduction

In this paper we study the many-server N-system shown in Fig. 1, with Poisson arrivals and exponential service times, under the first come first served and assign to the longest idle server policy (FCFS–ALIS), as the number of servers becomes large. Before describing the model in detail, we will first discuss our motivation for studying this system.

Research supported in part by Israel Science Foundation Grants 711/09 and 286/13.

✉ Dongyuan Zhan
d.zhan@ucl.ac.uk

Gideon Weiss
gweiss@stat.haifa.ac.il

[1] School of Management, University College London, 1 Canada Square, London E14 5AA, UK

[2] Department of Statistics, The University of Haifa, 31905 Mount Carmel, Israel

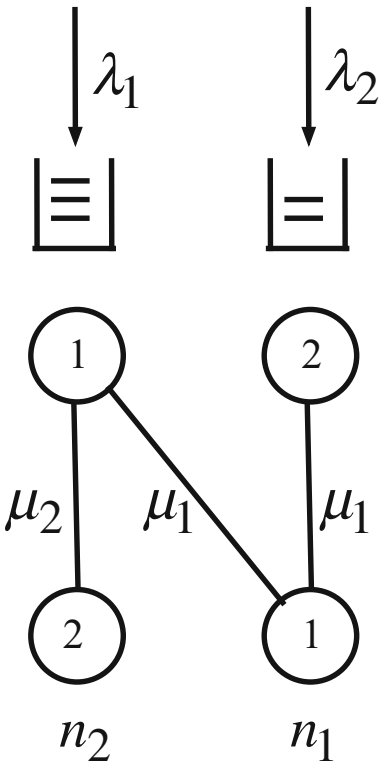

Fig. 1 The multi-server N-system

The N-system is one of the simplest special cases of skill-based routing in parallel server systems, as defined in [9,15] and further studied in [4,6,7,12–14,17,19,20,22,23]. The general model has customers of types $i = 1, \ldots, I$, servers of types $j = 1, \ldots, J$, and a bipartite compatibility graph G, where $(i, j) \in G$ if customer type i can be served by server type j. Arrivals are renewal with rate λ, where successive customer types are i.i.d. with probabilities α_i. There are a total of n servers, of which $n\theta_j$ are of type j, and service times are generally distributed with rates $\mu_{i,j}$. Assume the system is operated under the FCFS–ALIS policy, that is, servers take on the longest waiting compatible customer, and arriving customers are assigned to the longest idle compatible server. For this general system, necessary and sufficient conditions for stability (positive Harris recurrence for given λ), or for complete resource pooling (there exists a critical λ_0 such that the system is stable for $\lambda < \lambda_0$, and the queues of all customer types diverge for $\lambda > \lambda_0$) cannot be determined by the first moment information alone (as conjectured by an example of Foss and Chernova [9], which is further discussed in [16]). In particular, under FCFS–ALIS, calculation of the matching rates $r_{i,j}$, which are the long-term average fractions of services performed by servers of type j on customers of type i, in general, is intractable.

In the special case that service rates depend only on the server type, and not on the customer type, with Poisson arrivals and exponential service times, the system has a

product form stationary distribution, as given in [2]. In that case matching rates can be computed from the stationary distribution.

The following conjecture was made in [4]. If the system is stable and has complete resource pooling for given λ, n, and we let both become large together, the behavior of the system simplifies: there will exist β_j such that servers of type j perform a fraction β_j of the services, and the matching rates $r_{i,j}$ will converge to the rates for the FCFS infinite matching model with G, α, β, as calculated in [1] (see also [5]). The conjecture is based on the following heuristic argument: in steady state the times that each server becomes available form a stationary process which is only mildly correlated with the other servers, and so servers become available approximately as a superposition of almost independent stationary processes, which in the many-server limit becomes a Poisson process, and server types are then i.i.d. with probabilities β_j, while customer types arrive as an i.i.d. sequence with probabilities α_i. This corresponds exactly to the model of FCFS infinite matching. Under FCFS–ALIS it is also possible that while the system is stable, service by all the servers is not pooled. Instead it is decoupled: the bipartite compatibility graph breaks into two or more subgraphs, and when the system is operated under FCFS–ALIS the links connecting the subgraphs are only rarely used. The conjecture then is that under many-server scaling this decoupling is the same as in the FCFS infinite matching model, with the same matching rates.

In our current study of the many-server N-system we shall verify the conjectured many-server behavior for this simple parallel server system. To do so we start from the known stationary distribution of the N-system with many servers, as derived in [2], and study its behavior as $n \to \infty$. As it turns out, the product form stationary distribution, even for this simple case, is far from simple, and the derivations of limits, which use summations over server permutations and asymptotic expansions of various expressions, are quite laborious. We feel that this emphasizes the difficulty in verifying the conjectured behavior of the general system, which remains intractable at this time.

We mention that the N-system with just two servers has been the subject of several papers, including [3,10,11,19,20]. In this paper, our focus is on the N-system with many servers under FCFS–ALIS and its limiting behavior.

The rest of the paper is structured as follows. In Sect. 2 we describe the model, and in Sect. 3 we use some heuristic arguments to obtain a guess at the limiting behavior, where we distinguish between pooled and decoupled modes. In Sect. 4 we verify the heuristic guess and obtain the stationary behavior under many-server scaling. In Sect. 5 we illustrate our results with some numerical examples. To improve the readability of the paper we have put all the proofs for Sect. 4 in the Appendix.

2 The model

In our N-system, customers of types c_1 and c_2 arrive as independent Poisson streams, with rates λ_1, λ_2. There are skill-based parallel servers, n_1 servers of type s_1 which are flexible and can serve both types, and n_2 servers of type s_2 which can only serve type c_1 customers. In our notation, c_1 customers and s_1 servers are flexible, while c_2 customers and s_2 servers are inflexible. (s_2 servers cannot serve c_2 customers.) We assume service times are all independent exponential, with server-dependent rates.

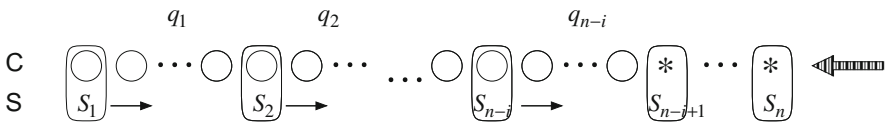

Fig. 2 State description under FCFS–ALIS

The service rate of an s_1 server is μ_1; the service rate of an s_2 server is μ_2. See Fig. 1. We let $\lambda = \lambda_1 + \lambda_2$, $n = n_1 + n_2$. The service policy is FCFS–ALIS.

The system is Markovian. In [2,3,21] the following state description for the skill-based parallel server systems under the FCFS–ALIS policy was used: imagine the customers arranged in a single queue by order of arrival, and servers are attached to the customers which they serve, and the remaining idle servers are arranged by increasing idle time in front of the queue; see Fig. 2. The state is then $\mathfrak{s} = (S_1, q_1, S_2, q_2, \ldots, S_{n-i}, q_{n-i}, S_{n-i+1}, \ldots, S_n)$, where $S_1, \ldots, S_n$ is a permutation of the n servers; the first $n - i$ servers are the ordered busy servers, and the last i servers are the ordered idle servers, and where q_j, $j = 1, \ldots, n - i$, are the queue lengths of the customers waiting for one of the servers $S_1, \ldots, S_j$, and skipped (could not be served) by servers $S_{j+1}, \ldots, S_n$. When service rates depend only on the servers, arrivals are Poisson, and services are exponential, this description is Markovian, as shown in [21]. The reason is as follows: given the permutation of servers, we know for each q_j exactly what types of customers may be present, and since those customers are in the order in which they arrived, the type of each of them is randomly distributed according to the initial frequencies of customer types, and independent of all others. Hence, each server with a queue in front will have to go through an independent sequence of trials as he scans the customers FCFS until finding a match, and the specific sequences of customer types in the queues are not relevant to the steady state of the scan. This yields Markovian transition probabilities.

For the special case of the N-system, in steady state, the following three random quantities are important: $i_1 = I_1(\mathfrak{s})$, the number of idle servers of type s_1, $i_2 = I_2(\mathfrak{s})$, the number of idle servers of type s_2, and $k = K(\mathfrak{s}) \geq 0$, the number of servers of type s_2 which follow the last server of type s_1 in the sequence $S_1, \ldots, S_n$. An incoming c_2 customer has to skip k s_2 servers and find the last s_1 server to be served. We let $i = I(\mathfrak{s})$ be the total number of idle servers in steady state. Because of the structure of the N-system and the FCFS–ALIS policy, the following properties hold for $i = 0, \ldots, n$ and $k = 0, \ldots, n_2$:

(i) There are no customers waiting for any server which precedes the last s_1 server in the permutation. In other words, for all $j < \min(n-k, n-i)$ we have $q_j = 0$. In particular, if there is an idle server of type s_1 (meaning $i > k$), then there are no waiting customers at all.
(ii) If there are any idle servers, then there are no type c_1 customers waiting for service; in other words, if $i > 0$, then all the waiting customers are of type c_2.
(iii) If there are no idle servers (all servers are busy), then only the last queue can contain type c_1 customers; in other words, if $i = 0$, then the last queue may contain customers of both types, but all the other waiting customers are of type c_2.

Denote

$$\alpha = \frac{\lambda_1}{\lambda}, \quad \theta = \frac{n_1}{n}, \quad \rho = \frac{\lambda}{n_1\mu_1 + n_2\mu_2}, \quad \delta = \frac{\lambda_2}{n_1\mu_1}, \quad r = \frac{\lambda}{n}.$$

Then a necessary and sufficient condition for stability is

$$\rho < 1, \quad \delta < 1.$$

Throughout the paper, we assume the above stability condition. For the stable system, define β as the long-term fraction of customers served by servers of type s_1, and $1-\beta$ the long-term fraction of customers served by servers of type s_2. Since type s_1 servers are the only ones that can serve type c_2 servers, we must have $\beta \geq 1-\alpha$, or, equivalently, $\alpha + \beta \geq 1$. The stable system under FCFS–ALIS may operate in two different modes: it may be that servers of both types share the service of customers of type c_1, in which case $\beta > 1-\alpha$ and we say that resource pooling occurs for large n, or it may be the case that servers of type s_1 serve almost exclusively only customers of type c_2, and almost all the service of customers of type c_1 is done by servers of type s_2, in which case $\beta \approx 1-\alpha$ for large n, and we say that the system is decoupled.

Using the results of [1,2] we can then write the exact stationary distribution of this system. We wish to show that, as the arrival rate and the number of servers increase, the system simplifies, and we get very precise many-server scaling limits, and in particular we find sharp conditions for pooled or decoupled modes of operation. We will investigate the behavior of the system when we fix the values of α, θ, ρ, and let $n \to \infty$. To be precise, we shall then have n, $n_1 = \lceil \theta n \rceil$, $n_2 = n - n_1, \lambda = \rho(\mu_1 n_1 + \mu_2 n_2)$, $\lambda_1 = \alpha\lambda$, $\lambda_2 = (1-\alpha)\lambda$, all of which go to infinity. Average processing times $1/\mu_1$, $1/\mu_2$ are fixed and not scaled.

3 Heuristic fluid calculations

In this section we use some heuristic arguments to guess at the fluid behavior of the many-server system. In particular, we calculate a guess for some key quantities. Using these quantities we give a heuristic description of how the system will behave under the FCFS–ALIS policy, in the many-server case, distinguishing between pooled and decoupled modes of operation. The main part of the paper, in Sect. 4, is the verification of these guesses.

We assume some fixed $\rho < 1$, $\delta < 1$ so that the system under FCFS–ALIS is stable. We then observe that under many-server scaling there will almost always be some idle servers available of both types and customers will almost never wait, so that they will enter service immediately upon arrival. At the same time, when a server completes a service there will almost never be any waiting customers, so, after almost every service completion, the server will experience some idle time. Because our policy is ALIS, when a server becomes idle, he always joins the end of a queue of idle servers. In a slight abuse of the notation, we reuse I_1, I_2 and K to denote, respectively, the stationary numbers of servers of type s_1, s_2 and the servers of type s_2 which follow the last server of type s_1 in $\mathfrak{s}$.

When the system is stationary, the sample path of each server will consist of a sequence of cycles, each of which consists of a single service period followed by an idle period (which can be equal to 0). We denote the generic idle periods between services by Y_1, Y_2. We can bound the values of $T_1,\ T_2$ as follows: servers of type s_2 can serve only customers of type c_1, some of which may also be served by servers of type s_1. Hence, the arrival rate per server is no larger than λ_1/n_2, and so the average interval between arrivals is no less than n_2/λ_1, and the average service time per arrival is $1/\mu_2$, hence $T_2 \geq n_2/\lambda_1 - 1/\mu_2$. Servers of type s_1 serve all customers of type c_2 and may in addition serve some customers of type c_1. Hence, the arrival rate per server is no less than λ_2/n_1, and so the average interval between arrivals is no larger than n_1/λ_2. The average service time per arrival is $1/\mu_1$; hence, $T_1 \leq n_1/\lambda_2 - 1/\mu_1$. Hence, we have found that the stationary expected idle time satisfies

$$T_1 = E(Y_1) \leq \frac{n_1}{\lambda_2} - \frac{1}{\mu_1}, \qquad T_2 = E(Y_2) \geq \frac{n_2}{\lambda_1} - \frac{1}{\mu_2}. \tag{1}$$

We now distinguish three cases for the values of the parameters:

$$\begin{aligned} \text{Case I} \quad & \frac{n_2}{\lambda_1} - \frac{1}{\mu_2} > \frac{n_1}{\lambda_2} - \frac{1}{\mu_1} \\ \text{Case II} \quad & \frac{n_2}{\lambda_1} - \frac{1}{\mu_2} < \frac{n_1}{\lambda_2} - \frac{1}{\mu_1} \\ \text{Case III} \quad & \frac{n_2}{\lambda_1} - \frac{1}{\mu_2} = \frac{n_1}{\lambda_2} - \frac{1}{\mu_1} \end{aligned}$$

Case I

In this case, by (1) we will have $T_2 > T_1$, and the system will decouple. The reasoning is as follows: because our policy is ALIS, each server, on completion of service, joins the end of the queue of idle servers, and his idle period consists of waiting until all the servers ahead of him who are of his type, as well as all the other servers that can serve customers who are compatible with him, are assigned to customers, and he is then assigned to the next compatible customer.

At the end of his idle period, a server of type s_i has been idle for Y_i, and he is then the longest idle server of his type. If we assume the idle times Y_i converge to their means T_i as the system becomes large, $i = 1, 2$, then since $T_2 > T_1$, we can say that most of the time the longest idle server will be of type s_2. Therefore almost all the arriving customers of type c_1 will be assigned to a server of type s_2, and so servers of type c_1 will serve almost only customers of type c_2.

This implies that in Case I the system under many-server scaling will behave like two separate $M/M/s$ queues. Because servers of type s_2 serve almost all customers of type c_1, and servers of type s_1 serve all customers of type c_2 and almost none of the customers of type c_1, we have, for large n,

$$\alpha + \beta \approx 1$$

and inequalities (1) will be close to equalities, and we will have (by Little's law)

$$E(I_1) = \lambda_2 T_1 \approx n_1 - \frac{\lambda_2}{\mu_1}, \quad E(I_2) = \lambda_1 T_2 \approx n_2 - \frac{\lambda_1}{\mu_2}.$$

We can also estimate the value of K, the location of the first type s_1 server. Since service completions of customers of type c_1 occur at rate λ_1 and almost all of those are served by type s_2, and service completions of customers of type c_1 occur at rate λ_1 and all of those and almost no others are served by type s_2, servers of type s_2 and s_1 join the end of the queue of idle servers at the ratio of λ_1/λ_2, so $(I_2 - K)/I_1 \approx \lambda_1/\lambda_2$ and

$$E(K) \approx E(I_2) - E(I_1)\frac{\lambda_1}{\lambda_2} = \lambda_1\big(T_2 - T_1\big) \approx \lambda_1 \left(\frac{n_2}{\lambda_1} - \frac{1}{\mu_2} - \frac{n_1}{\lambda_2} + \frac{1}{\mu_1}\right).$$

It is worthwhile to note that the condition of Case I that implies decomposition is not simply that $\delta > \rho$, which is equivalent to $\frac{\lambda_2}{n_1\mu_1} > \frac{\lambda_1}{n_2\mu_2}$ (the load of customers of type c_2 on servers of type s_1 is higher than the load of customers of type c_1 on servers of type s_2). In fact, under FCFS, servers of both types may share service of customers of type c_1 even when $\delta > \rho$. To explain, when $\delta > \rho$, under decoupled service, the load and therefore the busy time percentage of type s_2 servers is smaller than the load of type s_1 servers, but, if $\mu_1 < \mu_2$, the idle time of type s_2 servers (Y_2) could be shorter than that of type s_1 servers (Y_1). In that case, under FCFS the work of c_1 customers will be shared by both types of servers.

The stationary behavior of the decoupled system is described in Fig. 3. In this figure we have, from left to right, a section of busy servers of both types serving all the customers in the system, followed by a section of more recent queueing idle servers of mixed types, followed by a section of the oldest idle servers, all of which are of type s_2. Servers that complete service join the queue of idle servers at its left end. Arriving customers of type c_1 pick the oldest waiting server, which is of type c_2; arriving customers of type c_2 skip all the K servers of type s_2, and pick the oldest idle server of type s_1. Note that the idle servers of both types are mixed in the middle section, and $I_2 \neq I_1 + K$.

The exact limiting behavior under many-server scaling for Case I is derived in Sect. 4.4, where the heuristic calculations are verified. Our main results for Case I are:

- The probability that $K = 0$ converges to 0 as $n \to \infty$, and so every customer of type c_1 is served by a server of type s_2.

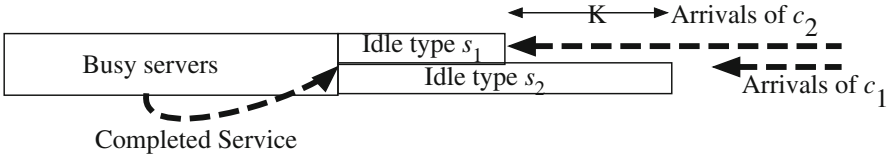

Fig. 3 FCFS–ALIS many-server system, queues of idle servers decoupled

- The two sets of servers and their customers behave like independent $M/M/n_1$ and $M/M/n_2$ queues.

Case II

In this case, we argue that $T_1 \to T_2$ as $n \to \infty$. Assume to the contrary that $T_1 > T_2$ as $n \to \infty$. Then, for large n, we should have that most of the time the longest idle server will be of type s_1. But s_1 servers can serve all customers, and so by ALIS s_1 servers will serve almost all the customers in the system, which is a contradiction. Now assume that $T_2 > T_1$ as $n \to \infty$. But in that case we already argued that the system will decouple and so the inequalities in (1) will hold as equalities, which, since we are in Case II, contradicts $T_2 > T_1$. Therefore, there is no decoupling in Case II, and we conclude that, for large n,

$$\frac{n_2}{\lambda_1} - \frac{1}{\mu_2} < T_2 \approx T_1 < \frac{n_1}{\lambda_2} - \frac{1}{\mu_1}.$$

Our first conclusion from $T_2 > \frac{n_2}{\lambda_1} - \frac{1}{\mu_2}$ is that servers of type s_2 do not serve all the customers of type c_1, so $1 - \beta < \alpha$, i.e., $\alpha + \beta > 1$, and from $T_1 < \frac{n_1}{\lambda_2} - \frac{1}{\mu_1}$ we conclude that servers of type s_1 serve some customers of type c_1 as well as customers of c_2 (again, $\beta > 1 - \alpha$).

The following is a heuristic description of the behavior of the system in Case II under many-server scaling. When n increases, the (random) number of idle servers becomes large, of order $O(n)$, and successive servers join the queue of idle servers at short intervals (of expected length $1/\lambda$, which is $O(1/n)$). They will spend a time of $O(1)$ to traverse the queue and will then reach the head of the queue of idle servers with short intervals between them. At this point they will need to wait for a compatible customer, and this waiting time does depend on the type of server, but because λ is large, once a server is at the head of the line his wait for a compatible customer will be short; hence, successive server arrivals to the idle queue are close to each other and so are their departures from the idle queue. So, as $n \to \infty$, not only does $T_1 = T_2$, but also the idle times, Y_1 and Y_2, have the same distribution, and K is of order $O(1)$. This heuristic description will be verified in Sect. 4.

We denote by T the presumed common value of T_1 and T_2. We now calculate the value of T. Let T be the average length of the idle time, common to all servers. The average cycle times will be $1/\mu_1 + T$ and $1/\mu_2 + T$. We defined β as the long-run fraction of services performed by s_1 servers, with $1 - \beta$ services by type s_2. The cycle rate of one type s_1 server is $1/(1/\mu_1 + T)$; hence, the processing rate of all type s_1 servers is $n_1/(1/\mu_1 + T)$, which should equal $\lambda\beta$. Similarly, the flow rate out of all type s_2 servers should equal $\lambda(1 - \beta)$. That is,

$$\lambda\beta = n_1/(1/\mu_1 + T), \quad \lambda(1 - \beta) = n_2/(1/\mu_2 + T).$$

Now we solve for T and β to obtain

$$\beta = \frac{n_1}{\lambda}\frac{1}{1/\mu_1 + T}, \quad 1 - \beta = \frac{n_2}{\lambda}\frac{1}{1/\mu_2 + T} \tag{2}$$

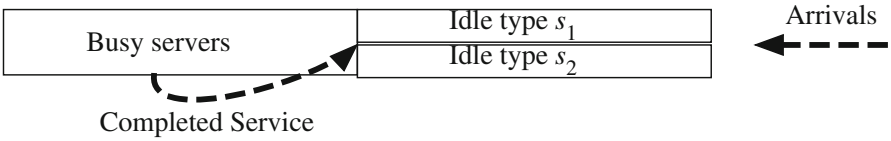

Fig. 4 FCFS–ALIS many-server system, queues of idle servers pooled

and a quadratic equation for T:

$$g(T) = \lambda\mu_1\mu_2 T^2 + \big(\lambda(\mu_1+\mu_2) - (n_1+n_2)\mu_1\mu_2\big)T + \lambda - n_1\mu_1 - n_2\mu_2 = 0.$$

Here $g(0) < 0$ because $\rho < 1$, so the equation has one positive and one negative root. Solving for positive T we get

$$\begin{aligned} T &= \frac{1}{2}\left(\frac{n}{\lambda} - \frac{1}{\mu_1} - \frac{1}{\mu_2} + \sqrt{\frac{n^2}{\lambda^2} + 2\,\frac{n_1-n_2}{\lambda}\left(\frac{1}{\mu_1}-\frac{1}{\mu_2}\right) + \left(\frac{1}{\mu_1}-\frac{1}{\mu_2}\right)^2}\right) \\ &= \frac{1}{2}\left(\frac{1}{\rho(\theta\mu_1+(1-\theta)\mu_2)} - \frac{1}{\mu_1} - \frac{1}{\mu_2}\right. \\ &\quad \left. + \sqrt{\frac{1}{\rho^2(\theta\mu_1+(1-\theta)\mu_2)^2} + \frac{4\theta-2}{\rho(\theta\mu_1+(1-\theta)\mu_2)}\left(\frac{1}{\mu_1}-\frac{1}{\mu_2}\right) + \left(\frac{1}{\mu_1}-\frac{1}{\mu_2}\right)^2}\right). \end{aligned} \tag{3}$$

Note: for the case of $\mu_1 = \mu_2 = \mu$ we get $T = \frac{1-\rho}{\rho}\frac{1}{\mu}$.

From T and Little's law we can obtain m_i, the approximate average number of idle servers in pool i, $i = 1, 2$:

$$m_1 = T\lambda\beta = \frac{Tn_1}{T + 1/\mu_1}, \quad m_2 = T\lambda(1-\beta) = \frac{Tn_2}{T + 1/\mu_2}. \tag{4}$$

When $T_1 = T_2$, servers are pooled. Servers share the load, and both types of customers receive similar levels of service. The pooled behavior of the system for FCFS–ALIS under many-server scaling is our main interest in this paper. Figure 4 shows the analog of Fig. 3 for the pooled system. Note that the idle servers of both types are mixed, and $I_2 \neq I_1$.

Case III

This case lies on the boundary of the other two cases. As a sanity check, on the one hand, we see that setting $T_1 = \frac{n_2}{\lambda_1} - \frac{1}{\mu_2}$ and $T_2 = \frac{n_2}{\lambda_1} - \frac{1}{\mu_2}$ would correspond to the values for Case I, and result in $T_1 = T_2$. On the other hand, considering the equation (2) for Case II, if we substitute

$$\beta = \frac{n_1}{\lambda}\frac{1}{1/\mu_1 + T} = \frac{n_1}{\lambda}\frac{1}{1/\mu_1 + T_1} = \frac{n_1}{\lambda}\frac{1}{1/\mu_1 + n_1/\lambda_2 - 1/\mu_1} = \frac{\lambda_2}{\lambda} = 1-\alpha,$$

$$1-\beta = \frac{n_2}{\lambda}\frac{1}{1/\mu_2 + T} = \frac{n_2}{\lambda}\frac{1}{1/\mu_1 + T_2} = \frac{n_2}{\lambda}\frac{1}{1/\mu_2 + n_2/\lambda_1 - 1/\mu_2} = \frac{\lambda_1}{\lambda} = \alpha,$$

therefore, $\alpha + \beta = 1$.

4 Many-server limit of the stationary distribution

In this section, we keep the stability assumption $\rho < 1$, $\delta < 1$ and derive the many-server limit from the exact stationary distributions.

4.1 Exact stationary distributions

We first obtain the stationary distribution for each state $\mathfrak{s}$. We note that the stationary probabilities depend mainly on the values of k, i_1, i_2. Let $\mu(S_j)$ denote the service rate of the server at position j.

Theorem 1 *The stationary distribution of the state $\mathfrak{s}$ of the FCFS–ALIS many-server N-system is given by*

$$\pi(\mathfrak{s}) = \begin{cases} B \prod_{l=1}^{n-i_1-i_2} \left(\sum_{j=1}^{l} \mu(S_j) \right)^{-1} \left(\frac{1}{\lambda} \right)^{i_1+i_2-k} \left(\frac{1}{\lambda_1} \right)^{k}, & \begin{array}{l} k = 0, \ldots, n_2, \\ i_1 = 1, \ldots, n_1, \\ i_2 = k, \ldots, n_2, \end{array} \\ B \prod_{l=1}^{n-k-1} \left(\sum_{j=1}^{l} \mu(S_j) \right)^{-1} \prod_{j=n-k}^{n-i_2} \frac{\lambda_2^{q_j}}{(\mu_1 n_1 + \mu_2 (j-n_1))^{q_j+1}} \left(\frac{1}{\lambda_1} \right)^{i_2}, & \begin{array}{l} k = 1, \ldots, n_2, \\ i_1 = 0, \\ i_2 = 1, \ldots, k, \end{array} \\ B \prod_{l=1}^{n-k-1} \left(\sum_{j=1}^{l} \mu(S_j) \right)^{-1} \prod_{j=n-k}^{n-1} \frac{\lambda_2^{q_j}}{(\mu_1 n_1 + \mu_2 (j-n_1))^{q_j+1}} \frac{\lambda^{q_n}}{(\mu_1 n_1 + \mu_2 n_2)^{q_n+1}}, & \begin{array}{l} k = 0, \ldots, n_2, \\ i_1 = i_2 = 0, \end{array} \end{cases} \tag{5}$$

where B is a normalizing constant.

Proof This follows for all three parts of (5) by utilizing properties (i),(ii),(iii) in Sect. 2 and substituting into Equation (2.1), Theorem 2.1, in [2]. □

Before we manipulate Eq. (5), we introduce a lemma to facilitate the calculation.

Lemma 1 *Letting $A_1, \ldots, A_m$ denote a permutation of m given positive real numbers $a_1, \ldots, a_m$, we have*

$$\sum_{(A_1,\ldots,A_m) \in \mathcal{P}(a_1,\ldots,a_m)} \prod_{l=1}^{m} \left(\sum_{j=1}^{l} A_j \right)^{-1} = \left(\prod_{l=1}^{m} a_l \right)^{-1}$$

where $\mathcal{P}(a_1, \ldots, a_m)$ denotes the set of all the permutations of $a_1, \ldots, a_m$.

Now we can get the joint stationary distribution of K, I_1, I_2. We denote by $\pi(k, i_1, i_2)$ the stationary probability of $K = k$, $I_1 = i_1$ and $I_2 = i_2$.

Theorem 2 *The steady-state joint distribution of* $K,\ I_1,\ I_2$ *is given by*

$$\pi(k,i_1,i_2)=\begin{cases} B_1\binom{n_1}{i_1}\binom{n_2}{i_2}\dfrac{i_1 i_2!(i_1+i_2-k-1)!}{(i_2-k)!}\mu_1^{i_1}\mu_2^{i_2}\left(\dfrac{1}{\lambda}\right)^{i_1+i_2}\left(\dfrac{\lambda}{\lambda_1}\right)^{k}, & \begin{array}{l}k=0,\ldots,n_2,\\ i_1=1,\ldots,n_1,\\ i_2=k,\ldots,n_2,\end{array}\\ B_1\dfrac{n_1\,n_2!}{(n_2-k)!}\mu_1\mu_2^{k}\displaystyle\prod_{j=n-k}^{n-i_2}\frac{1}{\mu_1 n_1+\mu_2(j-n_1)-\lambda_2}\left(\dfrac{1}{\lambda_1}\right)^{i_2}, & \begin{array}{l}k=1,\ldots,n_2,\\ i_1=0,\\ i_2=1,\ldots,k,\end{array}\\ B_1\dfrac{n_1\,n_2!}{(n_2-k)!}\mu_1\mu_2^{k}\displaystyle\prod_{j=n-k}^{n-1}\frac{1}{\mu_1 n_1+\mu_2(j-n_1)-\lambda_2}\frac{1}{\mu_1 n_1+\mu_2 n_2-\lambda}, & \begin{array}{l}k=0,\ldots,n_2,\\ i_1=i_2=0,\end{array}\end{cases} \tag{6}$$

where B_1 *is a normalizing constant.*

4.2 The distribution of (I_1, I_2) given K

In this section we obtain the asymptotic distribution of (I_1, I_2) conditional on $K = k$, as $n \to \infty$. We first show that, as $n \to \infty$, the probability of no idle servers of type s_1 goes to zero, and so the probability that customers need not wait goes to 1. Next we condition on $K = k$ and show $I_1/n \xrightarrow{p} f_1,\ I_2/n \xrightarrow{p} f_2$, where

$$f_1 = \frac{m_1}{n} = \frac{T\theta}{T + 1/\mu_1}, \quad f_2 = \frac{m_2}{n} = \frac{T(1-\theta)}{T + 1/\mu_2},$$

where T is given in (3). Finally, we condition on $K = k$ and show that the scaled and centered values of (I_1, I_2) converge in distribution to a bivariate normal distribution. Proofs of the following theorems can be found in the Appendix.

Theorem 3 *When* $n \to \infty$, *there exists an* $\epsilon > 0$ *such that*

$$P(I_1 = 0) = o\left(\exp(-\epsilon n)\right).$$

From this theorem we see that when $n \to \infty$, $P(I_1 > 0) \to 1$. Therefore, $P(K = k, I_1 > 0) \to P(K = k)$ for any $0 \le k \le I_2$. From Eq. (6), given $K = k$, the limiting stationary distribution as $n \to \infty$ is

$$\begin{aligned} &P(I_1 = i_1, I_2 = i_2 | K = k) \to P(I_1 = i_1, I_2 = i_2 | K = k, I_1 > 0)\\ &= B_1\binom{n_1}{i_1}\binom{n_2}{i_2} i_1(i_1+i_2-k-1)!\frac{i_2!}{(i_2-k)!}\mu_1^{i_1}\mu_2^{i_2}\lambda^{-i_1-i_2-k}\lambda_1^{-k}\frac{1}{P(K=k)}. \end{aligned}$$

Theorem 4 *Conditional on* $K = k$, $\left(\frac{I_1}{n},\ \frac{I_2}{n}\right)$ *converges to* $(f_1,\ f_2)$ *in probability for any* $k \ge 0$. *That is, for any* $\epsilon > 0$, *when* $n \to \infty$, *we have*

$$P\left(|I_1 - f_1 n| \ge \epsilon n \text{ or } |I_2 - f_2 n| \ge \epsilon n | K = k\right) \to 0.$$

After showing the fluid limit result, we are now ready to show the central limit result.

Theorem 5 *For any $k \geq 0$, when $n \to \infty$, we have*

$$\left(\frac{I_1 - f_1 n}{\sqrt{n}}, \frac{I_2 - f_2 n}{\sqrt{n}} \middle| K = k \right) \Rightarrow \mathcal{N}\left(0, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right), \tag{7}$$

where

$$\rho = \left(\frac{(\theta - f_1)(1 - \theta - f_2) f_1 f_2}{\left(\theta f_2 + f_1^2\right)\left((1-\theta) f_1 + f_2^2\right)} \right)^{\frac{1}{2}},$$

$$\sigma_1 = \left(\frac{(\theta - f_1) f_1 \left((1-\theta) f_1 + f_2^2\right)}{\theta f_2^2 + (1-\theta) f_1^2} \right)^{\frac{1}{2}},$$

$$\sigma_2 = \left(\frac{(1 - \theta - f_2) f_2 \left(\theta f_2 + f_1^2\right)}{\theta f_2^2 + (1-\theta) f_1^2} \right)^{\frac{1}{2}}.$$

Note that the above is consistent with the bivariate normal distribution stated in Sect. 3.

4.3 Case II: Pooled system

Now we consider Case II, where $\frac{n_2}{\lambda_1} - \frac{1}{\mu_2} < \frac{n_1}{\lambda_2} - \frac{1}{\mu_1}$. First we show the limit distribution of K, the location of the first type s_1 server.

Theorem 6 *In Case II, for any $k \geq 0$, as $n \to \infty$,*

$$P(K = k) \to \left(1 - \frac{1-\beta}{\alpha}\right)\left(\frac{1-\beta}{\alpha}\right)^k. \tag{8}$$

Theorem 6 shows that K converges in distribution to a geometric distribution in Case II, so $P(K < \infty) = 1$. Therefore, we can extend Theorems 4 and 5 into unconditional versions.

Theorem 7 *In Case II, as $n \to \infty$, K becomes independent of I_1 and I_2. $\left(\frac{I_1 - f_1 n}{\sqrt{n}}, \frac{I_2 - f_2 n}{\sqrt{n}}\right)$ converges in distribution to the bivariate normal distribution described in (10).*

Consider the special case when $\mu_1 = \mu_2 = \mu$. Then $\theta = \beta$, $f_1 = (1-\rho)\theta$ and $f_2 = (1-\rho)(1-\theta)$. When $n \to \infty$, $\left(\frac{I_1 - (1-\rho)n_1}{\sqrt{n}}, \frac{I_2 - (1-\rho)n_2}{\sqrt{n}}\right)$ converges in distribution to a bivariate normal distribution with mean $(0, 0)$, variance

$$\left(\rho\theta(1 - \rho(1-\theta)),\ \rho(1-\theta)(1-\rho\theta) \right),$$

and correlation

$$\frac{\rho\sqrt{\theta(1-\theta)}}{\sqrt{(1-\rho(1-\theta))(1-\rho\theta)}}.$$

The total idleness has mean of $(1-\rho)n$ and variance of

$$Var(I_1) + Var(I_2) + 2Cov(I_1, I_2) = \rho n.$$

4.4 Case I: Decoupling to two independent systems

We now assume $\frac{n_2}{\lambda_1} - \frac{1}{\mu_2} > \frac{n_1}{\lambda_2} - \frac{1}{\mu_1}$, where we find that under many-server scaling the system decouples into two independent $M/M/s$ service systems. We first show the following proposition:

Proposition 1 *In Case I, as $n \to \infty$, we have $P(\alpha I_1 \geq (1-\alpha) I_2) = o\left(\frac{1}{\sqrt{n}}\right)$.*

We next obtain the conditional distribution $K|(I_1, I_2)$.

Theorem 8 *Given $I_1 = i_1 n$, $I_2 = i_2 n$, where $i_1 \in (0, \theta)$, $i_2 \in (0, 1-\theta)$, and $i_2 > \frac{\alpha}{1-\alpha} i_1$, we have*

$$\left. \frac{K - \left(i_2 - \frac{\alpha}{1-\alpha} i_1\right) n}{\sqrt{n}} \right| (I_1 = i_1 n, I_2 = i_2 n) \Rightarrow \mathcal{N}\left(0, \frac{\alpha i_1}{(1-\alpha)^2}\right), \text{ as } n \to \infty.$$

Therefore, given $(1-\alpha) I_2 > \alpha I_1$, $P(K = 0 | I_1, I_2) = o\left(\frac{1}{\sqrt{n}}\right)$. Now we have

$$P(K=0) < P(K=0|I_1, I_2) + P((1-\alpha) I_2 \leq \alpha I_1) = o\left(\frac{1}{\sqrt{n}}\right).$$

That means the number of type c_1 customers served by s_1 servers is no more than $o(\sqrt{n})$, which cannot affect the fluid scaled mean or the diffusion scaled variance of two independent decoupled systems.

Theorem 9 *In Case I, as $n \to \infty$,*

$$\left(\frac{I_1 - \left(n_1 - \frac{\lambda_2}{\mu_1}\right)}{\sqrt{n}}, \frac{I_2 - \left(n_2 - \frac{\lambda_1}{\mu_2}\right)}{\sqrt{n}} \right) \Rightarrow \mathcal{N}\left(0, \begin{bmatrix} \frac{\lambda_2}{n\mu_1} & 0 \\ 0 & \frac{\lambda_1}{n\mu_2} \end{bmatrix}\right). \tag{9}$$

This is exactly the many-server scaling limiting distribution of the number of idle servers in two independent $M/M/s$ queues, one of which has arrival rate λ_2, service rate μ_1, and n_1 servers; the other has arrival rate λ_1, service rate μ_2, and n_2 servers.

Furthermore, K will then consist of I_2 minus the idle servers of type s_2 which are mingled with the I_1 servers of type s_1. The following calculation obtains the mean and variance of K under many-server scaling. We denote by $I_{2,1}$ the number of idle servers of type s_2 that are mingled with the I_1 idle servers of type s_1. Since the type s_1 servers join the idle servers with rate λ_2 and type s_2 servers join the idle servers with rate λ_1, we have

$$I_{2,1} = \sum_{j=1}^{I_1} W_i,$$

where W_i are i.i.d. random variables independent of I_1, each of them having the distribution of the number of failures before the first success in a sequence of Bernoulli trials with probability of success $\frac{\lambda_2}{\lambda_1+\lambda_2}$. We have

$$\begin{aligned}
E(W_i) &= \frac{\lambda_1}{\lambda_2},\\
\mathrm{Var}(W_i) &= \frac{\lambda_1(\lambda_1+\lambda_2)}{\lambda_2^2},\\
E(I_{2,1}) &= E(I_1)\frac{\lambda_1}{\lambda_2} = \left(n_1 - \frac{\lambda_2}{\mu_1}\right)\frac{\lambda_1}{\lambda_2},\\
\mathrm{Var}(I_{2,1}) &= E(I_1)\frac{\lambda_1(\lambda_1+\lambda_2)}{\lambda_2^2} + \mathrm{Var}(I_1)\left(\frac{\lambda_1}{\lambda_2}\right)^2\\
&= \left(n_1 - \frac{\lambda_2}{\mu_1}\right)\frac{\lambda_1(\lambda_1+\lambda_2)}{\lambda_2^2} + \frac{\lambda_2}{\mu_1}\left(\frac{\lambda_1}{\lambda_2}\right)^2.
\end{aligned}$$

Furthermore, as $n \to \infty$, centered and scaled $I_{2,1}$ converges to a normal distribution, and is independent of I_2.

It now follows that centered and scaled K also converges to a normal distribution, and centered and scaled (I_1, I_2, K) converge to a multivariate normal distribution. The relevant parameters are

$$\begin{aligned}
E(K) &= E(I_2) - E(I_{2,1}) = n_2 - \frac{\lambda_1}{\mu_2} - \left(n_1 - \frac{\lambda_2}{\mu_1}\right)\frac{\lambda_1}{\lambda_2}\\
&= \lambda_1\left(\frac{n_2}{\lambda_1} - \frac{1}{\mu_2} - \frac{n_1}{\lambda_2} + \frac{1}{\mu_1}\right),\\
\mathrm{Var}(K) &= \mathrm{Var}(I_2) + \mathrm{Var}(I_{2,1}) = \frac{\lambda_1}{\mu_2} + \left(n_1 - \frac{\lambda_2}{\mu_1}\right)\frac{\lambda_1(\lambda_1+\lambda_2)}{\lambda_2^2} + \frac{\lambda_2}{\mu_1}\left(\frac{\lambda_1}{\lambda_2}\right)^2\\
&= n_1\frac{\lambda_1\lambda}{\lambda_2^2} + \lambda_1\left(\frac{1}{\mu_2} - \frac{1}{\mu_1}\right).
\end{aligned}$$

K is correlated with both I_1 and I_2:

$$\begin{aligned}
\mathrm{Cov}(I_2, K) &= \mathrm{Cov}(I_2, I_2 - I_{2,1}) = \mathrm{Var}(I_2),\\
\mathrm{Cov}(I_1, K) &= \mathrm{Cov}(I_1, I_2 - I_{2,1}) = \mathrm{Cov}(I_1, -I_{2,1}) = -\frac{\lambda_1}{\lambda_2}\mathrm{Var}(I_1).
\end{aligned}$$

4.5 Case III: Slowly decoupling as system becomes large

As $n \to \infty$, we have seen that when $\frac{n_2}{\lambda_1} - \frac{1}{\mu_2} < \frac{n_1}{\lambda_2} - \frac{1}{\mu_1}$ (Case II), then $\frac{K}{n} \to 0$ in probability, and in fact $K = O(1)$; when $\frac{n_2}{\lambda_1} - \frac{1}{\mu_2} > \frac{n_1}{\lambda_2} - \frac{1}{\mu_1}$ (Case I), then $\frac{K}{n} \to \frac{\lambda_1}{n}\left(\frac{n_2}{\lambda_1} - \frac{1}{\mu_2} - \frac{n_1}{\lambda_2} + \frac{1}{\mu_1}\right) > 0$ in probability, and in fact $K = O(n)$. We now examine Case III, where $\frac{n_2}{\lambda_1} - \frac{1}{\mu_2} = \frac{n_1}{\lambda_2} - \frac{1}{\mu_1}$. We will show that in this case, as n becomes large, with fluid scaling the queues decouple, but with diffusion scaling K has nontrivial behavior.

We first prove a monotonicity result on K as a function of α, which holds for all three cases, I, II, and III. To mark dependence on α we use the notation K_α.

Proposition 2 *Keep all the other parameters fixed and change α. If $\alpha_1 < \alpha_2$, then K_{α_1} stochastically dominates K_{α_2}.*

From the monotonicity and the previous statements for Cases I and II, we conclude:

Corollary 1 *In Case III, as $n \to \infty$, $\frac{K}{n} \to 0$ in probability.*

We can in fact derive more precise asymptotic results for I_1, I_2, K in case III. We note first that the result of Theorem 5 on the limiting distribution of $\left(\frac{I_1 - m_1}{\sqrt{n}}, \frac{I_2 - m_2}{\sqrt{n}} \middle| K = k\right)$ as $n \to \infty$, for any fixed k, is valid not just in Case II, but also in Cases I and III. In the following theorem we investigate the limit, for fixed k, as $n \to \infty$, of $\left(\frac{I_1 - m_1}{\sqrt{n}}, \frac{I_2 - m_2}{\sqrt{n}} \middle| K = kn\right)$.

Theorem 10 *For any $k \in \left[0, 1 - \theta - \left[\frac{r - \theta\mu_1}{\mu_2}\right]^+\right)$, as $n \to \infty$, we have*

$$\left(\frac{I_1 - f_{1,k}n}{\sqrt{n}}, \frac{I_2 - f_{2,k}n}{\sqrt{n}} \middle| K = kn\right) \Rightarrow N\left(0, \begin{bmatrix} \sigma_{1,k}^2 & \rho_k\sigma_{1,k}\sigma_{2,k} \\ \rho_k\sigma_{1,k}\sigma_{2,k} & \sigma_{2,k}^2 \end{bmatrix}\right), \quad (10)$$

where

$$\rho_k = \left(\frac{f_{1,k}(f_{2,k} - k)(\theta - f_{1,k})(1 - \theta - f_{2,k})}{(f_{1,k}^2 + (f_{2,k} - k)\theta)((f_{2,k} - k)^2 + f_{1,k}(1 - \theta - k))}\right)^{\frac{1}{2}},$$

$$\sigma_{1,k} = \left(\frac{(\theta - f_{1,k})f_{1,k}((f_{2,k} - k)^2 + f_{1,k}(1 - \theta - k))}{f_{1,k}^2(1 - \theta - k) + (f_{2,k} - k)^2\theta}\right)^{\frac{1}{2}},$$

$$\sigma_{2,k} = \left(\frac{(1 - \theta - f_{2,k})(f_{2,k} - k)(f_{1,k}^2 + (f_{2,k} - k)\theta)}{f_{1,k}^2(1 - \theta - k) + (f_{2,k} - k)^2\theta}\right)^{\frac{1}{2}},$$

where $f_{1,k} = \frac{T\theta}{T + 1/\mu_1}$, $f_{2,k} = \frac{T(1-\theta-k)}{T + 1/\mu_2} + k$, and $T > 0$ solves

$$\frac{n_1}{\lambda}\frac{1}{1/\mu_1 + T} + \frac{n_2 - kn}{\lambda}\frac{1}{1/\mu_2 + T} = 1.$$

Note that $f_{i,0}$ equals f_i, defined in Sect. 4.2, for $i = 1, 2$. So when $k = 0$, Theorem 10 agrees with Theorem 5. We can now use these results to obtain the centered and scaled limiting behavior of K in Case III.

Theorem 11 *In Case III, as $n \to \infty$, $\frac{K}{\sqrt{n}}$ converges to a half truncated normal distribution with density function*

$$\sqrt{\frac{2}{\sigma_K^2 \pi}} \exp\left(-\frac{x^2}{2\sigma_K^2}\right), \forall x \geq 0,$$

where $\sigma_K^2 = \alpha\left(\frac{\lambda}{n}\left(\frac{1}{\mu_2} - \frac{1}{\mu_1}\right) + \frac{\theta}{(1-\alpha)^2}\right)$.

The result of Theorem 11 in combination with Theorem 10 should in principle allow us to obtain the joint distribution of (I_1, I_2). Its centered and scaled limit is, however, not a bivariate normal distribution, and too messy to write down. Theorem 11 directly implies that $P(K = 0) \to 0$ as $n \to \infty$. That means the proportion of type c_1 customers who are served by type s_1 servers goes to 0. Therefore, we can obtain the following fluid limit result:

Corollary 2 *In Case III,*

$$\lim_{n\to\infty} \frac{I_1 - \left(n_1 - \frac{\lambda_2}{\mu_1}\right)}{n} \to 0, \lim_{n\to\infty} \frac{I_2 - \left(n_2 - \frac{\lambda_1}{\mu_2}\right)}{n} \to 0,$$

which is the same as in Case I.

4.6 Comparison to the bipartite FCFS infinite matching model

The infinite matching model was defined and studied in [1,5,8] and is as follows: there are a set of customer types $\mathcal{C} = \{c_1, \ldots, c_I\}$ and a probability vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_I)$, a set of server types $\mathcal{S} = \{s_1, \ldots, s_J\}$ and a probability vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)$, and a bipartite compatibility graph $\mathcal{G} \subseteq \mathcal{C} \times \mathcal{S}$. There are two infinite sequences $C^1, C^2, \ldots$ where C^m are i.i.d. drawn from $\mathcal{C}$ with probabilities $\boldsymbol{\alpha}$, and $S^1, S^2, \ldots$ where S^n are i.i.d. drawn from $\mathcal{S}$ with probabilities $\boldsymbol{\beta}$. The two sequences are matched according to the compatibility graph, using FCFS. That is, C^1 is matched to the earliest S^n in the server sequence that is compatible with it, and thereafter C^m is matched to the earliest S^n in the server sequence that is compatible with it, and that was not matched to one of the customers $C^1, \ldots, C^{m-1}$. This model is much simpler than a parallel servers queueing model; because there are no arrival times, no busy or idle servers (only a sequence of service types), and no processing times, only ordered customer types and ordered service types matched in the FCFS manner. This model is tractable: under a condition of complete resource pooling the system reaches a steady state, and in particular it is possible to calculate the matching rate for each compatible pair r_{s_j,c_i}, the frequency of matches that happen between server type s_j and customer type c_i.

In the special case of the infinite matching model corresponding to the N-system, there are an infinite sequence of customers of types c_1, c_2, where the customer types are

i.i.d., the type is c_1 with probability α and c_2 with probability $1-\alpha$, and an independent infinite sequence of servers of types s_1, s_2, where the server types are i.i.d., the type is s_1 with probability β and s_2 with probability $1-\beta$, and the compatibility graph $\mathcal{G}$ has arcs $\{(c_1, s_1), (c_1, s_2), (c_2, s_1)\}$. The condition for complete resource pooling is then $\alpha + \beta > 1$, corresponding to Case II in our queueing model. Based on the exact formula in [1], successive customers and servers are matched according to FCFS, with matching rates $r_{c_1,s_1} = \alpha + \beta - 1$, $r_{c_1,s_2} = 1-\beta$, $r_{c_2,s_1} = 1-\alpha$.

After n customers have arrived and been matched, there may be some unmatched s_2 servers skipped by the customers. We define K_n to be the number of unmatched s_2 servers before the first unmatched s_1 server after the first n customers have been matched. We can see that $(K_n)_{n=1}^{\infty}$ is a Markov chain. If $K_n = 0$, that means server S^{n+1} is of type s_1, and then a new customer C^{n+1} will be matched to S^{n+1} and will add a geometrically distributed number with parameter β to K_n. If $K_n > 0$, then a new customer C^{n+1} of type c_1 will reduce K_n by 1, and a new customer C^{n+1} of type c_2 will add a geometrically distributed number with parameter β to K_n. The steady-state distribution for this Markov chain is that $P(K_\infty = k) = \left(1 - \frac{1-\beta}{\alpha}\right)\left(\frac{1-\beta}{\alpha}\right)^k$, $k \geq 0$, which is exactly the limiting distribution of K in (6). This supports our intuition that when the large N-system is underloaded with resource pooling in Case II, the replenishment of idle servers of types s_1 and s_2 becomes i.i.d with probability β and $1-\beta$, respectively.

In the infinite matching model, if complete resource pooling fails then there is a subset of customer types whose frequency is larger or equal to the frequency of all the compatible server types. In that case the infinite matching model will not reach steady state. However, in such cases there will be a unique decomposition of the model, so that each component on its own is an infinite matching model with complete resource pooling. In the case of the N-model this will happen when $\alpha + \beta \leq 1$, and then the model will decouple to two subsystems, one consisting of customers and servers of types c_1, s_2, and the other of customers and servers of types c_2, s_1. This is exactly the same decomposition that we observe in Cases I and III.

5 Numerical examples

We test our results by investigating an N-system with $\lambda = 100$, $n_1 = n_2 = 100$, $\mu_1 = \mu_2 = 1$, $\rho = 0.5$. In this example $\beta = 0.5$, $\theta\rho(1-\rho+\theta\rho)n = (1-\theta)\rho(1-\rho+(1-\theta)\rho)n = 37.5$. We use the exact stationary distribution to verify this. We calculate the expectation and variance of the idle number in each pool exactly, listed in the following table. In this example $\beta = 0.5$. When $\alpha > 0.5$ (Case II), so the average number of idle servers in each pool is close to 50, with variance close to $\theta\rho(1-\rho+\theta\rho)n = (1-\theta)\rho(1-\rho+(1-\theta)\rho)n = 37.5$; when $\alpha < 0.5$ (Case I), resource pooling disappears, and s_1 servers seldom serve c_1 customers. The N-system operates like two separate queues: s_1 servers server c_2 customers, and s_2 servers serve c_1 customers. The utilization of the s_1 server pool is $\frac{(1-\alpha)\lambda}{n_1}$, and the utilization of the s_2 server pool is $\frac{\alpha\lambda}{n_2}$. When $\alpha = 0.4$, almost zero portion of services performed by s_1 servers are for c_1 customers, the number of idle s_1 servers can be approximated by

Table 1 The exact calculation

α	$E[I_1]$	$\mathrm{Var}[I_1]$	$E[I_2]$	$\mathrm{Var}[I_2]$
0.8	49.838	37.705	50.162	37.381
0.7	49.648	38.078	50.352	37.374
0.6	49.179	39.215	50.821	37.572
0.55	48.606	40.871	51.395	38.082
0.5	47.333	44.883	52.667	39.549
0.4	39.981	59.821	60.019	39.785

a normal distribution with mean $n_1 - (1-\alpha)\lambda = 40$ and variance $(1-\alpha)\lambda = 60$, whereas the number of idle s_2 servers can be approximated by a normal distribution with mean $n_2 - \alpha\lambda = 60$ and variance $\alpha\lambda = 40$; when $\alpha = 0.5$ (Case III), we can see that the means are somewhat close to the fluid prediction 50, whereas we do not have analytic approximation for the variances (Table 1).

Acknowledgements We are grateful to Ivo Adan for helpful discussion of this paper. We thank the anonymous reviewer and the associate editor for their constructive comments, which helped us improve the manuscript. The review team noticed that analyzing only Case II left major gaps in the original version, which resulted in the addition of the analysis of Cases I and III.

A Appendix: Proofs for Sect. 4.1

Proof of Lemma 1 We prove this lemma by induction. Define the left-hand side as C_m.

$$C_2 = \frac{1}{a_1(a_1+a_2)} + \frac{1}{a_2(a_1+a_2)} = \frac{a_1+a_2}{a_1a_2(a_1+a_2)} = \frac{1}{a_1a_2}.$$

$$\begin{aligned} C_m &= \sum_{(A_1,\ldots,A_m)\in\mathcal{P}(a_1,\ldots,a_m)} \prod_{l=1}^{m} \left(\sum_{j=1}^{l} A_j\right)^{-1} \\ &= \frac{1}{\sum_{l=1}^{m} a_l} \sum_{p=1}^{m} \sum_{(A_1,\ldots,A_{m-1})\in\mathcal{P}(a_j:j\neq p)} \prod_{l=1}^{m-1} \left(\sum_{j=1}^{l} A_j\right)^{-1} \\ &= \frac{1}{\sum_{l=1}^{m} a_l} \sum_{p=1}^{m} \left(\prod_{j\neq p} a_j\right)^{-1} = \frac{1}{\sum_{l=1}^{m} a_l} \frac{\sum_{p=1}^{m} a_p}{\prod_{j=1}^{m} a_j} = \left(\prod_{l=1}^{m} a_l\right)^{-1}. \end{aligned}$$

□

Proof of Theorem 2 Summation over the geometric terms $q_j = 0, \ldots, \infty$ in (5) gives

$$\sum_{q_1,\ldots,q_{n-i}} \pi(\mathfrak{s}) = \begin{cases} B \prod_{l=1}^{n-i_1-i_2} \left(\sum_{j=1}^{l} \mu(S_j)\right)^{-1} \left(\frac{1}{\lambda}\right)^{i_1+i_2-k} \left(\frac{1}{\lambda_1}\right)^{k}, & k = 0, \ldots, n_2, \\ & i_1 = 0, \ldots, n_1 \\ & i_2 = k, \ldots, n_2, \\ B \prod_{l=1}^{n-k-1} \left(\sum_{j=1}^{l} \mu(S_j)\right)^{-1} \prod_{j=n-k}^{n-i_2} \frac{1}{\mu_1 n_1 + \mu_2(j-n_1) - \lambda_2} \left(\frac{1}{\lambda_1}\right)^{i_2}, & k = 1, \ldots, n_2, \\ & i_1 = 0, \\ & i_2 = 1, \ldots, k, \\ B \prod_{l=1}^{n-k-1} \left(\sum_{j=1}^{l} \mu(S_j)\right)^{-1} \prod_{j=n-k}^{n-1} \frac{1}{\mu_1 n_1 + \mu_2(j-n_1) - \lambda_2} \frac{1}{\mu_1 n_1 + \mu_2 n_2 - \lambda}, & k = 0, \ldots, n_2, \\ & i_1 = i_2 = 0. \end{cases}$$

Next we see that in this expression, permutations of $S_1, \ldots, S_n$ with the same (k, i_1, i_2) have a similar structure. We now sum over all the permutations of the appropriate S_j, $1 \le j \le n - \max\{k+1, i_1+i_2\}$. By Lemma 1 we obtain

$$\begin{cases} B\mu_1^{i_1-n_1}\mu_2^{i_2-n_2} \left(\frac{1}{\lambda}\right)^{i_1+i_2-k} \left(\frac{1}{\lambda_1}\right)^{k}, & k = 0, \ldots, n_2, \\ & i_1 = 1, \ldots, n_1, \\ & i_2 = k, \ldots, n_2, \\ B\mu_1^{1-n_1}\mu_2^{k-n_2} \prod_{j=n-k}^{n-i_2} \frac{1}{\mu_1 n_1 + \mu_2(j-n_1) - \lambda_2} \left(\frac{1}{\lambda_1}\right)^{i_2}, & k = 1, \ldots, n_2, \\ & i_1 = 0, \\ & i_2 = 1, \ldots, k, \\ B\mu_1^{1-n_1}\mu_2^{k-n_2} \prod_{j=n-k}^{n-1} \frac{1}{\mu_1 n_1 + \mu_2(j-n_1) - \lambda_2} \frac{1}{\mu_1 n_1 + \mu_2 n_2 - \lambda}, & k = 0, \ldots, n_2, \\ & i_1 = i_2 = 0. \end{cases} \tag{11}$$

Each permutation of the remaining servers, S_j, $n - \max\{k+1, i_1+i_2\} < j \le n$ has the same stationary probability. It remains to count the number of permutations. When $i_1 = 0$ we have $i_2 \le k$. For each permutation we choose 1 type s_1 server and k out of n_2 type s_2 servers to form the last $k+1$ servers. The number of permutations is

$$n_1 \binom{n_2}{k} k! = \frac{n_1\, n_2!}{(n_2-k)!}.$$

When $i_1 > 0$, we have $i_2 \ge k$. For each permutation, we choose i_1 out of n_1 type s_1 servers and i_2 out of n_2 type s_2 servers. We then choose 1 from the i_1 idle servers of type s_1, and k from the i_2 idle servers of type s_2 to obtain the last $k+1$ servers. The number of permutations is

$$\binom{n_1}{i_1}\binom{n_2}{i_2} i_1 \binom{i_2}{k} (i_1+i_2-k-1)!k! = \binom{n_1}{i_1}\binom{n_2}{i_2} \frac{i_1 i_2!(i_1+i_2-k-1)!}{(i_2-k)!}.$$

Multiplying the terms in (11) by the appropriate number of permutations and defining $B_1 = B\mu_1^{-n_1}\mu_2^{-n_2}$ gives (6). □

B Appendix: Proofs for Sect. 4.2

Proof of Theorem 3 We prove the theorem in three steps:

(i) We show that

$$P\left(I_1=0\right)\sim B_1\frac{1}{1-\delta}\times\begin{cases}\sqrt{2\pi n_2}\exp\left(n_2\left(-\log\kappa+\kappa-1\right)\right), & 0<\kappa<1,\\ \sqrt{2\pi n_2}/2, & \kappa=1,\\ \frac{1-(1-\alpha)\rho}{1-\rho}+\frac{1}{\kappa-1}, & \kappa>1,\end{cases}$$

where $\kappa=\frac{\lambda_1}{\mu_2 n_2}$. Note that $-\log\kappa+\kappa-1\geq 0$.

(ii) We show that

$$\begin{aligned}&P(I_1=\lceil m_1\rceil, I_2=\lceil m_2\rceil, K=0)\\ &\sim B_1\left(\frac{2\pi\beta n_1 n_2}{(n_1-m_1)(n_2-m_2)m_2}\right)^{1/2}\exp\left[-n_1\left(\log\left(1-\frac{m_1}{n_1}\right)+\frac{m_1}{n_1}\right)\right]\\ &\quad\exp\left[-n_2\left(\log\left(1-\frac{m_2}{n_2}\right)+\frac{m_2}{n_2}\right)\right],\end{aligned}$$

where $\sim$ means the ratio of the two sides converges to 1 when $n\to\infty$, m_1 and m_2 are defined in (4). Note that the definition in (4) does not require a specific case. And for all cases, we have $\frac{m_1}{m_2}=\frac{\beta}{1-\beta}$.

(iii) We show that, as $n\to\infty$,

$$\frac{P\left(I_1=0\right)}{P(I_1=\lceil m_1\rceil, I_2=\lceil m_2\rceil, K=0)}=o(\exp(-\epsilon n)),$$

for some $\epsilon>0$, which proves the theorem.

The details of the proofs of these three steps are as follows:

Proof of (i):

First we calculate

$$\begin{aligned}&P(I_1=0,\ I_2=0)=\sum_{k=0}^{n_2}\pi(k,0,0)\\ &=\sum_{k=0}^{n_2}B_1\frac{n_1\,n_2!}{(n_2-k)!}\mu_1\mu_2^k\prod_{j=n-k}^{n-1}\frac{1}{\mu_1 n_1+\mu_2(j-n_1)-\lambda_2}\,\frac{1}{\mu_1 n_1+\mu_2 n_2-\lambda}.\end{aligned}$$

We use induction to calculate

$$U_m=\sum_{k=m}^{n_2}\frac{\mu_2^k}{(n_2-k)!}\prod_{j=n-k}^{n-1}\frac{1}{\mu_1 n_1+\mu_2(j-n_1)-\lambda_2}$$

from $m=n_2$ to $m=1$. When $m=n_2$,

$$U_{n_2}=\frac{\mu_2^{n_2}}{(n_2-n_2)!}\frac{1}{\mu_1 n_1-\lambda_2}\prod_{j=n-n_2+1}^{n-1}\frac{1}{\mu_1 n_1+\mu_2(j-n_1)-\lambda_2}.$$

Suppose

$$U_{m+1} = \frac{\mu_2^{m+1}}{(n_2-m-1)!}\frac{1}{\mu_1 n_1-\lambda_2}\prod_{j=n-m}^{n-1}\frac{1}{\mu_1 n_1+\mu_2(j-n_1)-\lambda_2},$$

then

$$\begin{aligned}U_m &= \frac{\mu_2^{m+1}}{(n_2-m-1)!}\frac{1}{\mu_1 n_1-\lambda_2}\prod_{j=n-m}^{n-1}\frac{1}{\mu_1 n_1+\mu_2(j-n_1)-\lambda_2}\\ &\quad+\frac{\mu_2^{m}}{(n_2-m)!}\prod_{j=n-m}^{n-1}\frac{1}{\mu_1 n_1+\mu_2(j-n_1)-\lambda_2}\\ &= \frac{\mu_2^{m}}{(n_2-m)!}\prod_{j=n-m}^{n-1}\frac{1}{\mu_1 n_1+\mu_2(j-n_1)-\lambda_2}\left(\frac{\mu_2(n_2-m)}{\mu_1 n_1-\lambda_2}+1\right)\\ &= \frac{\mu_2^{m}}{(n_2-m)!}\frac{1}{\mu_1 n_1-\lambda_2}\prod_{j=n+1-m}^{n-1}\frac{1}{\mu_1 n_1+\mu_2(j-n_1)-\lambda_2}.\end{aligned}$$

Therefore, the induction is valid and we have

$$\begin{aligned}U_1 &= \frac{\mu_2}{(n_2-1)!}\frac{1}{\mu_1 n_1-\lambda_2}.\\ P(I_1=0,\ I_2=0) &= U_1 B_1 n_1 n_2!\frac{\mu_1}{\mu_1 n_1+\mu_2 n_2-\lambda}+\pi(0,0,0)\\ &= B_1\frac{\mu_2 n_2}{\mu_1 n_1-\lambda_2}\ \frac{\mu_1 n_1}{\mu_1 n_1+\mu_2 n_2-\lambda}+B_1\frac{\mu_1 n_1}{\mu_1 n_1+\mu_2 n_2-\lambda}\\ &= B_1\frac{\mu_1 n_1}{\mu_1 n_1-\lambda_2}\ \frac{\mu_1 n_1+\mu_2 n_2-\lambda_2}{\mu_1 n_1+\mu_2 n_2-\lambda}\\ &= B_1\frac{1}{1-\delta}\ \frac{1-(1-\alpha)\rho}{1-\rho}.\end{aligned}$$

Next we calculate

$$P(I_1=0,\ I_2>0)=\sum_{k=1}^{n_2}\sum_{i_2=1}^{k}\pi(k,0,i_2)=\sum_{i_2=1}^{n_2}\sum_{k=i_2}^{n_2}\pi(k,0,i_2).$$

Similar to the induction calculating U_m above, we can obtain

$$\begin{aligned}\sum_{k=i_2}^{n_2}\pi(k,0,i_2) &= B_1\left(\frac{1}{\lambda_1}\right)^{i_2} n_1\mu_1 n_2!\sum_{k=i_2}^{n_2}\frac{\mu_2^k}{(n_2-k)!}\prod_{j=n-k}^{n-i_2}\frac{1}{\mu_1 n_1+\mu_2(j-n_1)-\lambda_2}\\ &= B_1\left(\frac{1}{\lambda_1}\right)^{i_2} n_1\mu_1\, n_2!\frac{\mu_2^{i_2}}{(n_2-i_2)!}\frac{1}{\mu_1 n_1-\lambda_2}\\ &= B_1\frac{1}{1-\delta}\left(\frac{\mu_2}{\lambda_1}\right)^{i_2}\frac{n_2!}{(n_2-i_2)!}.\end{aligned}$$

Therefore,

$$\begin{aligned}P(I_1=0,\ I_2>0) &= B_1\frac{1}{1-\delta}n_2!\sum_{i_2=1}^{n_2}\left(\frac{\lambda_1}{\mu_2}\right)^{-i_2}\frac{1}{(n_2-i_2)!}\\ &= B_1\frac{1}{1-\delta}n_2!\left(\frac{\lambda_1}{\mu_2}\right)^{-n_2}\sum_{i_1'=0}^{n_2-1}\left(\frac{\lambda_1}{\mu_2}\right)^{i_2'}\frac{1}{i_2'!}\\ &= B_1\frac{1}{1-\delta}n_2!\left(\frac{\mu_2}{\lambda_1}\right)^{n_2}\exp\left(\frac{\lambda_1}{\mu_2}\right)P(X<n_2)\\ &= B_1\frac{1}{1-\delta}\frac{P(X<n_2)}{P(X=n_2)},\end{aligned}$$

where X is a Poisson random variable with parameter $\frac{\lambda_1}{\mu_2}$. Using Stirling's approximation,

$$\begin{aligned}P(X=n_2) &= \frac{1}{n_2!}\left(\frac{\lambda_1}{\mu_2}\right)^{n_2}\exp\left(-\frac{\lambda_1}{\mu_2}\right)\\ &\sim \frac{1}{\sqrt{2\pi n_2}}\left(\frac{\lambda_1}{\mu_2 n_2}\right)^{n_2}\exp\left(n_2-\frac{\lambda_1}{\mu_2}\right)\\ &= \frac{1}{\sqrt{2\pi n_2}}\exp\left(n_2\left(\log\left(\frac{\lambda_1}{\mu_2 n_2}\right)+1-\frac{\lambda_1}{\mu_2 n_2}\right)\right)\\ &= \frac{1}{\sqrt{2\pi n_2}}\exp\left(n_2\left(\log\kappa+1-\kappa\right)\right).\end{aligned}$$

Recall that $\kappa=\frac{\lambda_1}{\mu_2 n_2}$ and note that $\log\kappa+1-\kappa\le 0$. Note also that, when $n\to\infty$, X can be approximated by a normal distribution with mean $\frac{\lambda_1}{\mu_2}$ and variance $\frac{\lambda_1}{\mu_2}$. Next we analyze $\frac{P(X<n_2)}{P(X=n_2)}$ in three cases depending on κ.

- When $0<\kappa<1$, from the normal distribution approximation, when $n\to\infty$, $P(X<n_2)\to 1$. Therefore,

$$P\left(I_1=0,\ I_2>0\right)\sim B_1\frac{1}{1-\delta}\left(\sqrt{2\pi n_2}\exp\left(-n_2\left(\log\kappa+1-\kappa\right)\right)\right).$$

- When $\kappa = 1$, $-\log\kappa + \kappa - 1 = 0$. When $n \to \infty$, the normal distribution approximation gives $P(X < n_2) \to \frac{1}{2}$.

$$P\ (I_1 = 0,\ I_2 > 0) \sim B_1 \frac{1}{1-\delta}\frac{1}{2}\sqrt{2\pi n_2}.$$

- When $\kappa > 1$, when $n \to \infty$, the normal distribution approximation gives $P(X < n_2) \to 0$. We need more care to treat this case. For any $1 \le j \le n_2$,

$$\frac{P(X = n_2 - j)}{P(X = n_2)} = \frac{\left(\frac{\lambda_1}{\mu_2}\right)^{n_2-j}\frac{1}{(n_2-j)!}}{\left(\frac{\lambda_1}{\mu_2}\right)^{n_2}\frac{1}{n_2!}} = \frac{n_2!}{\kappa^j n_2^j (n_2-j)!} < \frac{1}{\kappa^j}.$$

Therefore,

$$\frac{P(X < n_2)}{P(X = n_2)} \le \sum_{j=1}^{n_2}\frac{1}{\kappa^j} < \frac{1}{\kappa - 1}.$$

In fact, for any fixed j, when $n \to \infty$,

$$\frac{P(X = n_2 - j)}{P(X = n_2)} \to \frac{1}{\kappa^j}.$$

For any $\epsilon > 0$, let $J = \lceil \frac{-\log\epsilon}{\log\kappa} \rceil$. We have $\epsilon \ge \kappa^{-J}$. There exists an N such that, when $n > N$, for any $1 \le j \le J$,

$$\frac{P(X = n_2 - j)}{P(X = n_2)} - \frac{1}{\kappa^j} > -\frac{\epsilon}{J}.$$

Therefore,

$$\frac{P(X < n_2)}{P(X = n_2)} > \sum_{j=1}^{J}\frac{1}{\kappa^j} - \epsilon = \frac{1-\kappa^{-J}}{\kappa - 1} - \epsilon \ge \frac{1}{\kappa - 1} - \frac{\kappa\epsilon}{\kappa - 1}.$$

Therefore, when $n \to \infty$,

$$\frac{P(X < n_2)}{P(X = n_2)} \to \frac{1}{\kappa - 1}.$$

We have

$$P\ (I_1 = 0,\ I_2 > 0) \sim B_1 \frac{1}{1-\delta}\frac{1}{\kappa - 1}.$$

In summary, when $\kappa \leq 1$, $P(I_1 = 0,\ I_2 = 0)$ is negligible compared with $P(I_1 = 0,\ I_2 > 0)$ when $n \to \infty$. We have

$$P(I_1 = 0) \sim B_1 \frac{1}{1-\delta} \times \begin{cases} \sqrt{2\pi n_2}\exp\left(n_2(-\log\kappa + \kappa - 1)\right), & 0 < \kappa < 1, \\ \sqrt{2\pi n_2}/2, & \kappa = 1, \\ \frac{1-(1-\alpha)\rho}{1-\rho} + \frac{1}{\kappa - 1}, & \kappa > 1. \end{cases}$$

Proof of (ii):

From Eq. (6) we have

$$\begin{aligned}
&P(I_1 = \lceil m_1 \rceil, I_2 = \lceil m_2 \rceil, K = 0) \\
&= B_1 \binom{n_1}{\lceil m_1 \rceil}\binom{n_2}{\lceil m_2 \rceil}\lceil m_1 \rceil(\lceil m_1 \rceil + \lceil m_2 \rceil - 1)!\mu_1^{\lceil m_1 \rceil}\mu_2^{\lceil m_2 \rceil}\left(\frac{1}{\lambda}\right)^{\lceil m_1 \rceil + \lceil m_2 \rceil} \\
&> \frac{B_1}{m_1^2 m_2 (m_1 + m_2)^2 \mu_1 \mu_2}\binom{n_1}{m_1}\binom{n_2}{m_2} m_1(m_1 + m_2 - 1)!\mu_1^{m_1}\mu_2^{m_2}\left(\frac{1}{\lambda}\right)^{m_1+m_2} \\
&> \frac{B_1}{n^5 \mu_1 \mu_2}\binom{n_1}{m_1}\binom{n_2}{m_2} m_1(m_1 + m_2 - 1)!\mu_1^{m_1}\mu_2^{m_2}\left(\frac{1}{\lambda}\right)^{m_1+m_2} \\
&\sim \frac{B_1}{n^5 \mu_1 \mu_2}\frac{m_1}{m_1 + m_2}\frac{n_1! n_2!}{(n_1 - m_1)! m_1! (n_2 - m_2)! m_2!}(m_1 + m_2)!\mu_1^{m_1}\mu_2^{m_2}\lambda^{-m_1-m_2} \\
&\sim \frac{B_1}{n^5 \mu_1 \mu_2}\left(\frac{2\pi m_1 n_1 n_2}{(m_1 + m_2)(n_1 - m_1)(n_2 - m_2)m_2}\right)^{1/2}\frac{n_1^{n_1} n_2^{n_2}}{(n_1 - m_1)^{n_1 - m_1} m_1^{m_1} (n_2 - m_2)^{n_2 - m_2} m_2^{m_2}} \\
&\quad \times \left(\frac{m_1 + m_2}{e}\right)^{m_1+m_2}\left(\frac{\mu_1}{\lambda}\right)^{m_1}\left(\frac{\mu_2}{\lambda}\right)^{m_2} \\
&= \frac{B_1}{n^5 \mu_1 \mu_2}\left(\frac{2\pi m_1 n_1 n_2}{(m_1 + m_2)(n_1 - m_1)(n_2 - m_2)m_2}\right)^{1/2} \\
&\quad \times \left(\frac{n_1}{n_1 - m_1}\right)^{n_1}\left(\frac{n_2}{n_2 - m_2}\right)^{n_2}\left(\frac{m_1 + m_2}{m_1}\right)^{m_1}\left(\frac{m_1 + m_2}{m_2}\right)^{m_2} \\
&\quad \times \exp(-m_1 - m_2)\left(\frac{\mu_1(n_1 - m_1)}{\lambda}\right)^{m_1}\left(\frac{\mu_2(n_2 - m_2)}{\lambda}\right)^{m_2} \\
&= \frac{B_1}{n^5 \mu_1 \mu_2}\left(\frac{2\pi\beta n_1 n_2}{(n_1 - m_1)(n_2 - m_2)m_2}\right)^{1/2}\left(\frac{n_1}{n_1 - m_1}\right)^{n_1}\left(\frac{n_2}{n_2 - m_2}\right)^{n_2}\exp(-m_1 - m_2) \\
&= \frac{B_1}{n^5 \mu_1 \mu_2}\left(\frac{2\pi\beta n_1 n_2}{(n_1 - m_1)(n_2 - m_2)m_2}\right)^{1/2}\exp\left(-n_1\left(\log\left(1 - \frac{m_1}{n_1}\right) + \frac{m_1}{n_1}\right)\right) \\
&\quad \times \exp\left(-n_2\left(\log\left(1 - \frac{m_2}{n_2}\right) + \frac{m_2}{n_2}\right)\right).
\end{aligned}$$

The second equality is due to $\frac{m_1}{m_1+m_2} = \beta$, $\frac{m_2}{m_1+m_2} = 1 - \beta$, $\frac{\mu_1(n_1 - m_1)}{\lambda} = \beta$, $\frac{\mu_2(n_2 - m_2)}{\lambda} = 1 - \beta$.

Proof of (iii):

Since $\log(1 - x) + x < 0$ when $0 < x < 1$, we have

$$\log\left(1 - \frac{m_1}{n_1}\right) + \frac{m_1}{n_1} < 0 \text{ and } \log\left(1 - \frac{m_2}{n_2}\right) + \frac{m_2}{n_2} < 0.$$

When $n \to \infty$, note that $\left(\frac{2\pi\beta n_1 n_2}{(n_1-m_1)(n_2-m_2)m_2}\right)^{1/2}$ is of the order of $n^{-1/2}$. Therefore, $P(I_1 = \lceil m_1 \rceil, I_2 = \lceil m_2 \rceil, K = 0)/B_1$ increases exponentially. When $\kappa > 1$, $P(I_1 = 0)/B_1$ converges to a constant; when $\kappa = 1$, $P(I_1 = 0)/B_1$ increases in the order of $\sqrt{n}$. Therefore, when $n \to \infty$ and $\kappa \geq 1$,

$$\frac{P\,(I_1 = 0)}{P(I_1 = \lceil m_1 \rceil, I_2 = \lceil m_2 \rceil, K = 0)} = o(\exp(-\epsilon n)),$$

for some $\epsilon > 0$. When $\kappa < 1$,

$$\begin{aligned}\frac{P\,(I_1 = 0)}{P(I_1 = \lceil m_1 \rceil, I_2 = \lceil m_2 \rceil, K = 0)} &\sim \left(\frac{(n_1 - m_1)(n_2 - m_2)m_2}{\beta n_1 (1-\delta)^2}\right)^{1/2} \\ &\times \exp\left(n_1\left(\log\left(1 - \frac{m_1}{n_1}\right) + \frac{m_1}{n_1}\right)\right) \\ &\times \exp\left(n_2\left(\log\left(1 - \frac{m_2}{n_2}\right) + \frac{m_2}{n_2} - \log\kappa + \kappa - 1\right)\right).\end{aligned}$$

We have that

$$\begin{aligned}\log\left(1 - \frac{m_2}{n_2}\right) + \frac{m_2}{n_2} - \log\kappa + \kappa - 1 &= \log\left(\frac{n_2 - m_2}{n_2}\frac{\mu_2 n_2}{\lambda_1}\right) + \frac{\lambda_1 - (n_2 - m_2)\mu_2}{n_2\mu_2} \\ &= \log\left(\frac{(n_2 - m_2)\mu_2}{\lambda_1}\right) + \frac{\lambda_1 - (n_2 - m_2)\mu_2}{\lambda_1}\kappa \\ &< \log\left(\frac{1-\beta}{\alpha}\right) + \frac{\alpha - (1-\beta)}{\alpha} \\ &= \log\left(1 - \frac{\alpha + \beta - 1}{\alpha}\right) + \frac{\alpha + \beta - 1}{\alpha},\end{aligned}$$

which is nonpositive no matter whether $\alpha + \beta$ is larger than, equal to, or small than 1. Therefore, when $n \to \infty$,

$$\frac{P\,(I_1 = 0)}{P(I_1 = \lceil m_1 \rceil, I_2 = \lceil m_2 \rceil, K = 0)} = o(\exp(-\epsilon n)),$$

for some $\epsilon > 0$. This completes the proof. □

Proof of Theorem 4 First we show that the weak convergence is valid given $K = 0$. Then we show that the same holds when $K = k$, for any fixed k. When $K = 0$, we prove the convergence in probability in two steps:

(i) We show that for all states $|I_1 - m_1| \geq \epsilon n$ or $|I_2 - m_2| \geq \epsilon n$, the conditional probability is dominated by a bounded constant multiple of the conditional probability of some point on the boundary of the rectangle $|I_1 - m_1| \leq \epsilon n \times |I_2 - m_2| \leq \epsilon n$.
(ii) When $n \to \infty$, we approximate the conditional probability of the points in the rectangle $|I_1 - m_1| \leq \epsilon n \times |I_2 - m_2| \leq \epsilon n$. We then show that the probability

of points on the boundary is negligible compared with the conditional probability at $(\lceil m_1 \rceil, \lceil m_2 \rceil)$.

Proof of (i):

$$\begin{aligned}
&P(I_1 = i_1, I_2 = i_2 | K = 0) \\
&= B_2 \binom{n_1}{i_1} \binom{n_2}{i_2} i_1 (i_1 + i_2 - 1)! \mu_1^{i_1} \mu_2^{i_2} \lambda^{-i_1 - i_2} \\
&= B_2 \frac{n_1! n_2!}{(n_1 - i_1)!(i_1 - 1)!(n_2 - i_2)! i_2!} (i_1 + i_2 - 1)! \left(\frac{\mu_1}{\lambda}\right)^{i_1} \left(\frac{\mu_2}{\lambda}\right)^{i_2},
\end{aligned}$$

where $B_2 = B_1 / P(K = 0)$.

$$\frac{P(I_1 = i_1 + 1, I_2 = i_2 | K = 0)}{P(I_1 = i_1, I_2 = i_2 | K = 0)} = \frac{(i_1 + i_2)(n_1 - i_1)\mu_1}{i_1 \lambda} = \beta \frac{n_1 - i_1}{n_1 - m_1} \frac{i_1 + i_2}{i_1}.$$
$$\frac{P(I_1 = i_1, I_2 = i_2 + 1 | K = 0)}{P(I_1 = i_1, I_2 = i_2 | K = 0)} = \frac{(i_1 + i_2)(n_2 - i_2)\mu_2}{(i_2 + 1)\lambda} = (1 - \beta) \frac{n_2 - i_2}{n_2 - m_2} \frac{i_1 + i_2}{i_2 + 1}.$$

We look at several cases:

- when $i_1 \le m_1$ and $(1 - \beta) i_1 < \beta i_2$, we have $\frac{i_1 + i_2}{i_1} > \frac{1}{\beta}$. Therefore, $\frac{P(I_1 = i_1 + 1, I_2 = i_2 | K = 0)}{P(I_1 = i_1, I_2 = i_2 | K = 0)} > 1$;
- when $i_2 \le m_2$ and $(1 - \beta) i_1 > \beta i_2 + 1$, we have $\frac{i_1 + i_2}{i_2 + 1} > \frac{1}{1 - \beta}$. Therefore, $\frac{P(I_1 = i_1, I_2 = i_2 + 1 | K = 0)}{P(I_1 = i_1, I_2 = i_2 | K = 0)} > 1$;
- when $i_1 > m_1$, $i_2 > m_2$ and $(1 - \beta) i_1 \ge \beta i_2$, we have $\frac{i_1 + i_2}{i_1} \le \frac{1}{\beta}$. Therefore, $\frac{P(I_1 = i_1 + 1, I_2 = i_2 | K = 0)}{P(I_1 = i_1, I_2 = i_2 | K = 0)} < 1$;
- when $i_1 > m_1, i_2 > m_2$ and $(1 - \beta) i_1 \le \beta i_2 + 1$, we have $\frac{i_1 + i_2}{i_2 + 1} \le \frac{1}{1 - \beta}$. Therefore, $\frac{P(I_1 = i_1, I_2 = i_2 + 1 | K = 0)}{P(I_1 = i_1, I_2 = i_2 | K = 0)} < 1$;
- when $\beta i_2 \le (1 - \beta) i_1 \le \beta i_2 + 1$, $i_1 \le m_1 - \epsilon n$ and $i_2 \le m_2 - \epsilon n$, as long as $\frac{n_2 - i_2}{n_2 - m_2} \frac{i_2}{i_2 + 1} > 1$, we have $\frac{P(I_1 = i_1, I_2 = i_2 + 1 | K = 0)}{P(I_1 = i_1, I_2 = i_2 | K = 0)} > 1$. When n is large, this requires

$$i_2 > i_2^* = \frac{1 - \theta - f_2}{f_2}.$$

As long as $\frac{n_1 - i_1}{n_1 - m_1} \frac{i_1 - 1}{i_1} > 1$, we have $\frac{P(I_1 = i_1 + 1, I_2 = i_2 | K = 0)}{P(I_1 = i_1, I_2 = i_2 | K = 0)} > 1$. When n is large, this requires

$$i_1 > i_1^* = \frac{\theta}{f_1}.$$

For all $i_1 > i_1^*$ or $i_2 > i_2^*$, we can move the state to a neighbor state with larger steady-state probability, as shown in Fig. 5.

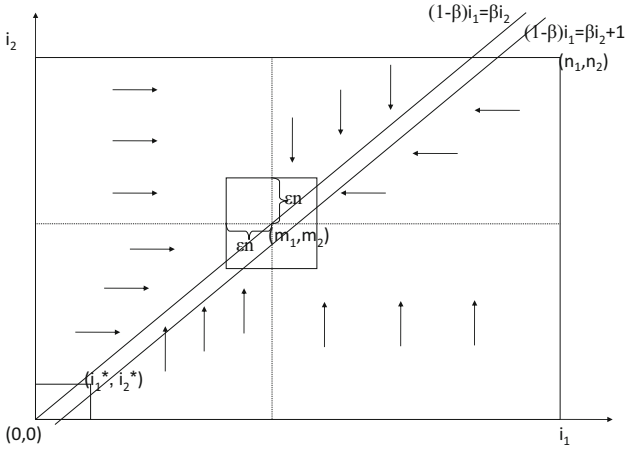

Fig. 5 The dominance of steady-state probability

Eventually the movement stops at the boundary which is ϵn away from (m_1, m_2). Therefore, the probability of any state (i_1, i_2) satisfying $i_1 > i_1^*$ or $i_2 > i_2^*$ would be dominated by the probability of some point at the boundary.

For any (i_1, i_2) satisfying $i_1 \le i_1^*$ and $i_2 \le i_2^*$, since

$$\frac{P(I_1 = i_1 + 1, I_2 = i_2 | K = 0)}{P(I_1 = i_1, I_2 = i_2 | K = 0)} > \beta \text{ and } \frac{P(I_1 = i_1, I_2 = i_2 + 1 | K = 0)}{P(I_1 = i_1, I_2 = i_2 | K = 0)} > 1 - \beta,$$

we have

$$P(I_1 = i_1, I_2 = i_2 | K = 0) < \frac{1}{\beta^{i_1^*+1}(1-\beta)^{i_2^*+1}} P(I_1 = i_1^* + 1, I_2 = i_2^* + 1 | K = 0)$$

and $P(I_1 = i_1^* + 1, I_2 = i_2^* + 1 | K = 0)$ is dominated by the probability of some point at the boundary.

Proof of (ii):

When $i_1 \in [m_1 - \epsilon n, m_1 + \epsilon n]$ and $i_2 \in [m_2 - \epsilon n, m_2 + \epsilon n]$, and n grows large, we can use Stirling's approximation.

$$\begin{aligned}
&P(I_1 = i_1, I_2 = i_2 | K = 0) \\
&= B_2 \binom{n_1}{i_1} \binom{n_2}{i_2} i_1 (i_1 + i_2 - 1)! \mu_1^{i_1} \mu_2^{i_2} \lambda^{-i_1 - i_2} \\
&= B_2 \frac{i_1}{i_1 + i_2} \frac{n_1! n_2!}{(n_1 - i_1)! i_1! (n_2 - i_2)! i_2!} (i_1 + i_2)! \left(\frac{\mu_1}{\lambda}\right)^{i_1} \left(\frac{\mu_2}{\lambda}\right)^{i_2} \\
&\sim B_3 \left(\frac{i_1}{(i_1 + i_2)(n_1 - i_1)(n_2 - i_2) i_2} \right)^{1/2} \frac{(i_1 + i_2)^{i_1 + i_2} \exp(-i_1 - i_2)}{(n_1 - i_1)^{n_1 - i_1} i_1^{i_1} (n_2 - i_2)^{n_2 - i_2} i_2^{i_2}} \left(\frac{\mu_1}{\lambda}\right)^{i_1} \left(\frac{\mu_2}{\lambda}\right)^{i_2} \\
&= B_3 \left(\frac{i_1}{(i_1 + i_2)(n_1 - i_1)(n_2 - i_2) i_2} \right)^{1/2} \exp\big((i_1 + i_2) \log(i_1 + i_2)
\end{aligned}$$

$$
\begin{aligned}
&-(n_1-i_1)\log(n_1-i_1)-i_1\log(i_1)\\
&-(n_2-i_2)\log(n_2-i_2)-i_2\log(i_2)+i_1\log\left(\frac{\mu_1}{\lambda}\right)+i_2\log\left(\frac{\mu_2}{\lambda}\right)-i_1-i_2\big)\\
=\ & B_3\left(\frac{i_1}{(i_1+i_2)(n_1-i_1)(n_2-i_2)i_2}\right)^{1/2}\exp\big(n\big((x_1+x_2)\log(x_1+x_2)\\
&-(\theta-x_1)\log(\theta-x_1)-x_1\log(x_1)\\
&-(1-\theta-x_2)\log(1-\theta-x_2)-x_2\log(x_2)+x_1\log\left(\frac{\mu_1}{r}\right)\\
&+x_2\log\left(\frac{\mu_2}{r}\right)-x_1-x_2-\log(n)\big)\big),
\end{aligned}
$$

where $B_3 = B_2 n_1! n_2! (2\pi)^{-\frac{3}{2}} \mathrm{e}^n$, $x_1 = \frac{i_1}{n}$, $x_2 = \frac{i_2}{n}$. We have $x_1 \in [f_1-\epsilon, f_1+\epsilon]$ and $x_2 \in [f_2-\epsilon, f_2+\epsilon]$. We define

$$
\begin{aligned}
F(x_1,x_2) &= (x_1+x_2)\log(x_1+x_2)-(\theta-x_1)\log(\theta-x_1)\\
&-(1-\theta-x_2)\log(1-\theta-x_2)\\
&+x_1(\log\mu_1-\log r-\log x_1)+x_2(\log\mu_2-\log r-\log x_2)-x_1-x_2.
\end{aligned}
$$

The first-order derivatives on x_1 and x_2 are

$$
\begin{aligned}
\frac{\partial F}{\partial x_1} &= \log(x_1+x_2)+\log(\theta-x_1)-\log(x_1)-\log\frac{\mu_1}{r}=0,\\
\frac{\partial F}{\partial x_2} &= \log(x_1+x_2)+\log(1-\theta-x_2)-\log(x_2)-\log\frac{\mu_2}{r}=0.
\end{aligned}
$$

Solving the first-order conditions gives

$$
x_1 = f_1, \qquad x_2 = f_2.
$$

Consider the second-order derivatives:

$$
\begin{aligned}
\frac{\partial^2 F}{\partial x_1^2} &= -\frac{1}{\theta-x_1}-\frac{1}{x_1}+\frac{1}{x_1+x_2}<0,\\
\frac{\partial^2 F}{\partial x_2^2} &= -\frac{1}{1-\theta-x_2}-\frac{1}{x_2}+\frac{1}{x_1+x_2}<0,\\
\frac{\partial^2 F}{\partial x_1\partial x_2} &= \frac{1}{x_1+x_2},\\
\frac{\partial^2 F}{\partial x_1^2}\frac{\partial^2 F}{\partial x_2^2}-\left(\frac{\partial^2 F}{\partial x_1\partial x_2}\right)^2 &= \frac{x_1^2(1-\theta)+x_2^2\theta}{x_1x_2(x_1+x_2)(\theta-x_1)(1-\theta-x_2)}>0.
\end{aligned}
$$

The Hessian matrix is negative definite. Therefore, $F(x_1, x_2)$ is strictly concave on $(0,\theta)\times(0,1-\theta)$ and reaches its unique global maximum at (f_1, f_2). The maximum of $F(x_1, x_2)$ on $[\delta, \theta-\delta]\times[\delta, 1-\theta-\delta]\backslash(f_1-\epsilon, f_1+\epsilon)\times(f_2-\epsilon, f_2+\epsilon)$ is on the boundary $\{(x_1,x_2)||x_1-f_1|=\epsilon, |x_2-f_2|=\epsilon\}$. Since the boundary is a compact set, the maximum is attainable, denoted by $F(f_1, f_2)-\eta$, where $\eta>0$.

Note that

$$\left(\frac{i_1}{(i_1+i_2)(n_1-i_1)(n_2-i_2)i_2}\right)^{1/2} = \left(\frac{x_1}{(x_1+x_2)(\theta-x_1)(1-\theta-x_2)x_2}\right)^{1/2} n^{-1}$$

changes slowly when x_1 and x_2 change, compared with $\exp(nF(x_1, x_2))$. We have

$$\frac{P(I_1 = i_1, I_2 = i_2|K=0)}{P(I_1 = \lceil m_1 \rceil, I_2 = \lceil m_2 \rceil|K=0)} \sim \exp(n(F(x_1, x_2) - F(f_1, f_2))) < \exp(\eta n).$$

Therefore,

$$\frac{\sum_{|i_1-m_1|>\epsilon n \text{ or } |i_2-m_2|>\epsilon n} P(I_1 = i_1, I_2 = i_2|K=0)}{P(I_1 = \lceil m_1 \rceil, I_2 = \lceil m_2 \rceil|K=0)} < \left(n_1 n_2 + \frac{(i_1^*+1)(i_2^*+1)}{\beta^{i_1^*+1}(1-\beta)^{i_2^*+1}}\right)\exp(\eta n).$$

It converges to 0 when $n \to \infty$.

When $K = k > 0$, and $n \to \infty$, similarly,

$$\begin{aligned}
&P(I_1 = i_1, I_2 = i_2|K=k)\\
&= B_1\binom{n_1}{i_1}\binom{n_2}{i_2} i_1(i_1+i_2-k-1)!\frac{i_2!}{(i_2-k)!}\mu_1^{i_1}\mu_2^{i_2}\lambda^{-i_1-i_2-k}\lambda_1^k \Big/ P(K=k)\\
&\quad \frac{P(I_1 = i_1+1, I_2 = i_2|K=k)}{P(I_1 = i_1, I_2 = i_2|K=k)}\\
&= \frac{(i_1+i_2-k)(n_1-i_1)\mu_1}{i_1\lambda} = \beta\frac{n_1-i_1}{n_1-m_1}\frac{i_1+i_2-k}{i_1}\\
&\quad \frac{P(I_1 = i_1, I_2 = i_2+1|K=k)}{P(I_1 = i_1, I_2 = i_2|K=k)}\\
&= \frac{(i_1+i_2-k)(n_2-i_2)\mu_2}{(i_2+1-k)\lambda} = (1-\beta)\frac{n_2-i_2}{n_2-m_2}\frac{i_1+i_2-k}{i_2+1-k}.
\end{aligned}$$

We can use a similar two-step argument to show that $(I_1/n,\ I_2/n)$ converges to $(f_1,\ f_2)$ in probability given $K = k$. □

Proof of Theorem 5 To obtain the asymptotic distribution of I_1, I_2 as $n \to \infty$, we need to consider, by Theorem 4, only values i_1, i_2 for which $(i_1 - m_1)/n \to 0$ and $(i_2 - m_2)/n \to 0$. We write $i_1 = m_1 + z_1\sqrt{n},\ i_2 = m_2 + z_2\sqrt{n}$, with $z_1/\sqrt{n} \to 0$, $z_2/\sqrt{n} \to 0$. Note that $m_1,\ m_2,\ n_1 - m_1,\ n_2 - m_2$ are of the same order of magnitude as n, n_1, n_2, and we only consider i_1, i_2 of the same order of magnitude.

$$
\begin{aligned}
&P(I_1 = i_1, I_2 = i_2 | K = 0) \\
&= B_2 \frac{i_1}{i_1 + i_2} \frac{n_1! n_2!}{(n_1 - i_1)! i_1! (n_2 - i_2)! i_2!} (i_1 + i_2)! \mu_1^{i_1} \mu_2^{i_2} \lambda^{-i_1 - i_2} \\
&\sim B_3 i_1^{-i_1} (n_1 - i_1)^{-(n_1 - i_1)} i_2^{-i_2} (n_2 - i_2)^{-(n_2 - i_2)} \left(\frac{i_1 + i_2}{e} \right)^{i_1 + i_2} \mu_1^{i_1} \mu_2^{i_2} \lambda^{-i_1 - i_2} \\
&\quad \times \left(\frac{i_1}{(i_1 + i_2) i_2 (n_1 - i_1)(n_2 - i_2)} \right)^{1/2},
\end{aligned}
$$

where the use of Stirling's approximation is justified for large n. Here $B_2 = B_1 / P(K = 0)$ and $B_3 = B_2 n_1! n_2! (2\pi)^{-\frac{3}{2}} \mathrm{e}^n$.

We clearly have

$$
\left(\frac{i_1}{(i_1 + i_2) i_2 (n_1 - i_1)(n_2 - i_2)} \right)^{1/2} \sim \left(\frac{m_1}{(m_1 + m_2) m_2 (n_1 - m_1)(n_2 - m_2)} \right)^{1/2},
$$

so we can treat that part as a constant. Consider

$$
i_1^{i_1} = (m_1 + z_1 \sqrt{n})^{m_1 + z_1 \sqrt{n}} = m_1^{m_1 + z_1 \sqrt{n}} \left(1 + \frac{z_1 \sqrt{n}}{m_1} \right)^{m_1 + z_1 \sqrt{n}}.
$$

Then from the Taylor expansion of the logarithm function, we have

$$
\begin{aligned}
&\log \left(\left(1 + \frac{z_1 \sqrt{n}}{m_1} \right)^{m_1 + z_1 \sqrt{n}} \right) = (m_1 + z_1 \sqrt{n}) \log \left(1 + \frac{z_1 \sqrt{n}}{m_1} \right) \\
&= (m_1 + z_1 \sqrt{n}) \left(\frac{z_1 \sqrt{n}}{m_1} - \frac{z_1^2 n}{2 m_1^2} + o\left(\frac{1}{n} \right) \right) = z_1 \sqrt{n} + \frac{z_1^2 n}{2 m_1} + o(1).
\end{aligned}
$$

Therefore,

$$
\log \left(i_1^{i_1} \right) \sim m_1 \log(m_1) + z_1 \sqrt{n} (\log(m_1) + 1) + \frac{z_1^2 n}{2 m_1}.
$$

Similar expansions are valid for i_2, $n_1 - i_1$, $n_2 - i_2$ and $i_1 + i_2$:

$$
\begin{aligned}
\log(i_2^{i_2}) &\sim m_2 \log(m_2) + z_2 \sqrt{n} (\log(m_2) + 1) + \frac{z_2^2 n}{2 m_2}. \\
\log((n_i - i_1)^{n_1 - i_1}) &\sim (n_1 - m_1) \log(n_1 - m_1) \\
&\quad - z_1 \sqrt{n} (\log(n_1 - m_1) + 1) + \frac{z_1^2 n}{2 (n_1 - m_1)}. \\
\log((n_2 - i_2)^{n_2 - i_2}) &\sim (n_2 - m_2) \log(n_2 - m_2) \\
&\quad - z_2 \sqrt{n} (\log(n_2 - m_2) + 1) + \frac{z_2^2 n}{2 (n_2 - m_2)}.
\end{aligned}
$$

$$\log((i_1+i_2)^{i_1+i_2}) \sim (m_1+m_2)\log(m_1+m_2) + (z_1+z_2)\sqrt{n}(\log(m_1+m_2)+1) + \frac{(z_1+z_2)^2 n}{2(m_1+m_2)}.$$

We now use the calculations in Sect. 3 to evaluate all the $\sqrt{n}$ coefficients. By (4) we have

$$\frac{(n_1-m_1)\mu_1}{\lambda m_1} = \frac{1}{\lambda T}, \quad \frac{(n_2-m_2)\mu_2}{\lambda m_2} = \frac{1}{\lambda T},$$
$$m_1+m_2 = T\left(\frac{n_1}{T+1/\mu_1}+\frac{n_2}{T+1/\mu_2}\right) = \lambda T.$$

Therefore, we have

$$\frac{(m_1+m_2)(n_1-m_1)\mu_1}{m_1\lambda} = 1, \quad \frac{(m_1+m_2)(n_2-m_2)\mu_2}{m_2\lambda} = 1,$$
$$\log\left(P(I_1=i_1, I_2=i_2|K=0)\right) \sim B_4 + \frac{(z_1+z_2)^2 n}{2(m_1+m_2)} - \frac{z_1^2 n n_1}{2(n_1-m_1)m_1} - \frac{z_2^2 n n_2}{2(n_2-m_2)m_2},$$

where $B_4 = \log\left(B_3\left(\frac{m_1}{(m_1+m_2)m_2(n_1-m_1)(n_2-m_2)}\right)^{1/2}\right) - n_1\log(n_1-m_1) - n_2\log(n_2-m_2) - m_1 - m_2$. Define

$$\rho = \left(\frac{(n_1-m_1)(n_2-m_2)m_1m_2}{(n_1m_2+m_1^2)(n_2m_1+m_2^2)}\right)^{\frac{1}{2}} = \left(\frac{(\theta-f_1)(1-\theta-f_2)f_1f_2}{(\theta f_2+f_1^2)((1-\theta)f_1+f_2^2)}\right)^{\frac{1}{2}},$$
$$\sigma_1 = \left(\frac{(n_1-m_1)m_1(n_2m_1+m_2^2)}{n_1m_2^2+n_2m_1^2}\right)^{\frac{1}{2}} = \left(\frac{(\theta-f_1)f_1((1-\theta)f_1+f_2^2)}{\theta f_2^2+(1-\theta)f_1^2}\right)^{\frac{1}{2}},$$
$$\sigma_2 = \left(\frac{(n_2-m_2)m_2(n_1m_2+m_1^2)}{n_1m_2^2+n_2m_1^2}\right)^{\frac{1}{2}} = \left(\frac{(1-\theta-f_2)f_2(\theta f_2+f_1^2)}{\theta f_2^2+(1-\theta)f_1^2}\right)^{\frac{1}{2}}.$$

We have

$$P(I_1=i_1, I_2=i_2|K=0) \sim \exp(B_4)\exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{z_1^2 n}{\sigma_1^2}+\frac{z_2^2 n}{\sigma_2^2}-\frac{2\rho z_1 z_2 n}{\sigma_1\sigma_2}\right)\right).$$

Therefore, $(\frac{I_1-m_1}{\sqrt{n}}, \frac{I_2-m_2}{\sqrt{n}})$ given $K=0$ converges in distribution as $n\to\infty$ to the bivariate normal distribution as stated in (10).

When $K = k > 0$, and $n \to \infty$, similarly,

$$\begin{aligned}
&P(I_1 = i_1, I_2 = i_2 | K = k) \\
&= B_1 \binom{n_1}{i_1}\binom{n_2}{i_2} i_1 (i_1 + i_2 - k - 1)! \frac{i_2!}{(i_2 - k)!} \mu_1^{i_1} \mu_2^{i_2} \lambda^{-i_1 - i_2 - k} \lambda_1^k \Big/ P(K = k) \\
&\sim B_1 \alpha^k \frac{i_1 i_2^k}{(i_1 + i_2)^{k+1}} \frac{n_1! n_2!}{(n_1 - i_1)! i_1! (n_2 - i_2)! i_2!} (i_1 + i_2)! \mu_1^{i_1} \mu_2^{i_2} \lambda^{-i_1 - i_2} \Big/ P(K = k).
\end{aligned}$$

We again write $i_1 = m_1 + z_1\sqrt{n}$, $i_2 = m_2 + z_2\sqrt{n}$, with $z_1/\sqrt{n} \to 0$, $z_2/\sqrt{n} \to 0$. We then have

$$\frac{i_1 i_2^k}{(i_1 + i_2)^{k+1}} \to \frac{m_1 m_2^k}{(m_1 + m_2)^{k+1}} = \beta(1 - \beta)^k.$$

We can now use the same approximation as for $k = 0$ to show that $\left(\frac{I_1 - m_1}{\sqrt{n}}, \frac{I_2 - m_2}{\sqrt{n}}\right)$ converges to the same bivariate normal distribution. □

C Appendix: Proofs for Sect. 4.3

Proof of Theorem 6 From (4),

$$\frac{f_2}{f_1 + f_2} = \frac{m_2}{m_1 + m_2} = \frac{T\lambda(1 - \beta)}{T\lambda\beta + T\lambda(1 - \beta)} = 1 - \beta.$$

Take a fixed arbitrary $\epsilon \in (0, \min\{f_1, f_2\})$. Fix $k > 0$. For any i_1, i_2 satisfying $|i_1/n - f_1| < \epsilon$, $|i_2/n - f_2| < \epsilon$ and $i_1 \geq 1$, from (6), noting $\frac{a+c}{b+c} \geq \frac{a}{b}$ for any $0 < a \leq b$ and $c > 0$, we have

$$\begin{aligned}
\frac{\pi(k, i_1, i_2)}{\pi(k - 1, i_1, i_2)} &= \frac{i_2 - k + 1}{i_1 + i_2 - k} \frac{1}{\alpha} \leq \frac{i_2 + 1}{i_1 + i_2} \frac{1}{\alpha} \leq \frac{(f_2 + \epsilon)n + 1}{i_1 + (f_2 + \epsilon)n} \frac{1}{\alpha} \\
&< \frac{(f_2 + \epsilon)n + 1}{(f_1 - \epsilon)n + (f_2 + \epsilon)n} \frac{1}{\alpha} \\
&= \frac{(f_2 + \epsilon)n + 1}{(f_1 + f_2)n} \frac{1}{\alpha} = \frac{f_2}{(f_1 + f_2)\alpha} \left(1 + \frac{\epsilon}{f_2} + \frac{1}{f_2 n}\right) = \frac{1 - \beta}{\alpha} \left(1 + \frac{\epsilon}{f_2} + \frac{1}{f_2 n}\right).
\end{aligned} \tag{12}$$

Therefore,

$$\pi(k, i_1, i_2) < \pi(0, i_1, i_2) \left(\frac{1 - \beta}{\alpha}\right)^k \left(1 + \frac{\epsilon}{f_2} + \frac{1}{f_2 n}\right)^k.$$

For fixed $k_0 > 0$,

$$P\left(K \geq k_0, I_1 = i_1, I_2 = i_2\right) < \pi(0, i_1, i_2) \frac{\left(\frac{1-\beta}{\alpha}\right)^{k_0} \left(1 + \frac{\epsilon}{f_2} + \frac{1}{f_2 n}\right)^{k_0}}{1 - \left(\frac{1-\beta}{\alpha}\right)\left(1 + \frac{\epsilon}{f_2} + \frac{1}{f_2 n}\right)}.$$

Note the above inequality is valid for any i_1, i_2 satisfying $|i_1/n - f_1| < \epsilon, |i_2/n - f_2| < \epsilon$. We have

$$\frac{P\left(K \geq k_0, |I_1/n - f_1| < \epsilon, |I_2 - f_2| < \epsilon\right)}{P\left(K = 0, |I_1/n - f_1| < \epsilon, |I_2 - f_2| < \epsilon\right)} < \frac{\left(\frac{1-\beta}{\alpha}\right)^{k_0} \left(1 + \frac{\epsilon}{f_2} + \frac{1}{f_2 n}\right)^{k_0}}{1 - \left(\frac{1-\beta}{\alpha}\right)\left(1 + \frac{\epsilon}{f_2} + \frac{1}{f_2 n}\right)}.$$

From Theorem 4, there exists an N_1 such that, when $n > N_1$,

$$P(|I_1/n - f_1| < \epsilon, |I_2/n - f_2| < \epsilon) > 1 - \epsilon.$$

Then we have,

$$\begin{aligned}
P(K \geq k_0) &< P\left(K \geq k_0 \,|\, |I_1/n - f_1| \geq \epsilon, |I_2/n - f_2| \geq \epsilon\right) \\
&\quad \times P\left(|I_1/n - f_1| < \epsilon, |I_2/n - f_2| < \epsilon\right) \\
&\quad + P\left(|I_1/n - f_1| \geq \epsilon, |I_2/n - f_2| \geq \epsilon\right) \\
&< P\left(K = 0 \,|\, |I_1/n - f_1| \geq \epsilon, |I_2/n - f_2| \geq \epsilon\right) \\
&\quad \frac{\left(\frac{1-\beta}{\alpha}\right)^{k_0} \left(1 + \frac{\epsilon}{f_2} + \frac{1}{f_2 n}\right)^{k_0}}{1 - \left(\frac{1-\beta}{\alpha}\right)\left(1 + \frac{\epsilon}{f_2} + \frac{1}{f_2 n}\right)} (1 - \epsilon) + \epsilon \\
&< \frac{\left(\frac{1-\beta}{\alpha}\right)^{k_0} \left(1 + \frac{\epsilon}{f_2} + \frac{1}{f_2 n}\right)^{k_0}}{1 - \left(\frac{1-\beta}{\alpha}\right)\left(1 + \frac{\epsilon}{f_2} + \frac{1}{f_2 n}\right)} (1 - \epsilon) + \epsilon.
\end{aligned}$$

This upper bound can be arbitrarily close to 0 when choosing ϵ, $n > N_1$, and k_0. Therefore, we have shown the tightness of K; that is,

$$\sum_{k=0}^{\infty} \lim_{n \to \infty} P\left(K = k\right) = 1. \tag{13}$$

Using

$$\begin{aligned}
P(K = k) &= P\left(K = k, |I_1/n - f_1| < \epsilon, |I_2/n - f_2| < \epsilon\right) \\
&\quad + P\left(K = k, |I_1/n - f_1| \geq \epsilon, |I_2/n - f_2| \geq \epsilon\right),
\end{aligned}$$

for fixed $k > 0$, when $n > N_1$, the ratio $\frac{P(K=k)}{P(K=k-1)}$ is lower bounded by

$$\frac{P\left(K=k, |I_1/n - f_1| < \epsilon, |I_2/n - f_2| < \epsilon\right)}{P\left(K=k-1, |I_1/n - f_1| < \epsilon, |I_2/n - f_2| < \epsilon\right) + \epsilon}$$

and upper bounded by

$$\frac{P\left(K=k, |I_1/n - f_1| < \epsilon, |I_2/n - f_2| < \epsilon\right) + \epsilon}{P\left(K=k-1, |I_1/n - f_1| < \epsilon, |I_2/n - f_2| < \epsilon\right)}.$$

For any i_1, i_2 satisfying $|i_1/n - f_1| < \epsilon$, $|i_2/n - f_2| < \epsilon$ and $i_1 \geq 1$, in addition to (12), we have the lower bound

$$\begin{aligned}\frac{\pi(k, i_1, i_2)}{\pi(k-1, i_1, i_2)} &= \frac{i_2 - k + 1}{i_1 + i_2 - k}\frac{1}{\alpha} \geq \frac{(f_2-\epsilon)n - k + 1}{i_1 + (f_2-\epsilon)n - k}\frac{1}{\alpha} \\ &> \frac{(f_2-\epsilon)n - k + 1}{(f_1+\epsilon)n + (f_2-\epsilon)n - k}\frac{1}{\alpha} = \frac{(f_2-\epsilon)n - k + 1}{(f_1+f_2)n - k}\frac{1}{\alpha}.\end{aligned}$$

Now we have

$$\frac{\pi(k, i_1, i_2)}{\pi(k-1, i_1, i_2)} \in \left[\frac{(f_2-\epsilon)n - k + 1}{(f_1+f_2)n - k}\frac{1}{\alpha}, \frac{(f_2+\epsilon)n + 1}{(f_1+f_2)n}\frac{1}{\alpha}\right].$$

Therefore,

$$\frac{\sum_{|i_1/n - f_1| < \epsilon, |i_2/n - f_2| < \epsilon} \pi(k, i_1, i_2)}{\sum_{|i_1/n - f_1| < \epsilon, |i_2/n - f_2| < \epsilon} \pi(k-1, i_1, i_2)} \in \left[\frac{(f_2-\epsilon)n - k + 1}{(f_1+f_2)n - k}\frac{1}{\alpha}, \frac{(f_2+\epsilon)n + 1}{(f_1+f_2)n}\frac{1}{\alpha}\right],$$

that is,

$$\begin{aligned}&\frac{P\left(K=k, |I_1/n - f_1| < \epsilon, |I_2/n - f_2| < \epsilon\right)}{P\left(K=k-1, |I_1/n - f_1| < \epsilon, |I_2/n - f_2| < \epsilon\right)} \\ &\in \left[\frac{(f_2-\epsilon)n - k + 1}{(f_1+f_2)n - k}\frac{1}{\alpha}, \frac{(f_2+\epsilon)n + 1}{(f_1+f_2)n}\frac{1}{\alpha}\right].\end{aligned} \tag{14}$$

For fixed k, as $n \to \infty$, the lower bound and the upper bound in (14) both converge to $\frac{1-\beta}{\alpha}$. Noting that ϵ can be arbitrarily close to 0, we have

$$\lim_{n\to\infty} \frac{P(K=k)}{P(K=k-1)} = \frac{1-\alpha}{\beta}.$$

This, together with the tightness (13), proves (8). □

Proof of Theorem 7 When $\frac{n_1}{\lambda_2} - \frac{1}{\mu_1} > \frac{n_2}{\lambda_1} - \frac{1}{\mu_2}$, the unscaled K converges to a geometric distribution. As we saw in Theorems 4 and 5, as $n \to \infty$, the distribution of the scaled deviations of I_1, I_2 conditional on the value of $K = k$ converges to a

normal distribution, with mean and variance that do not depend on k. We can now use the law of total probability and find N_0 large enough so that the unconditional probability distribution of the scaled I_1, I_2 is close to the specified normal distribution when $n > N_0$. One more step then shows that, as $n \to \infty$, the conditional distribution given K is the same, so we have the asymptotic independence. □

D Appendix: Proofs for Sect. 4.4

Proof of Proposition 1 Let $A_1(t)$ be the arrival stream of customers that are served eventually by servers of type s_1, and let $I_1(t)$ be, as defined above, the number of idle servers of type s_1. We now compare this to an $M/M/n_1$ system, with type s_1 servers, whose processing times are exponential with rate μ_1, and with arrival stream $\tilde{A}_1(t)$ which consists of all the arrivals of the stream $A_1(t)$ which are customers of type c_2, but excludes arrivals of type c_1. Clearly, $A_1(t) \geq \tilde{A}_1(t)$ a.s. Denote by $\tilde{I}_1(t)$ the number of idle servers in the $M/M/n_1$ system at time t. It then follows directly from Theorem 1 of Shanthikumar and Yao [18] that the stationary distributions of I_1 and $\tilde{I}_1$ satisfy $\tilde{I}_1 \geq_{ST} I_1$.

Define similarly an $M/M/n_2$ system with type s_2 servers, whose processing times are exponential with rate μ_2 and arrivals $\tilde{A}_2(t)$ of all the customers of type c_1. Then $A_2(t) \leq \tilde{A}_2(t)$ a.s. and, by the same argument, $\tilde{I}_2 \leq_{ST} I_2$.

As n becomes large, the numbers of idle servers in the two independent $M/M/N$ systems $(\tilde{I}_1(\infty), \tilde{I}_2(\infty))$ can be approximated by normal distributions with means

$$n_1 - \frac{\lambda_2}{\mu_1}, \quad n_2 - \frac{\lambda_1}{\mu_2}$$

and standard deviations $\sqrt{\frac{\lambda_2}{\mu_1}}$ and $\sqrt{\frac{\lambda_1}{\mu_2}}$, respectively. Since, in Case I,

$$c = \left(\frac{n_2}{\lambda_1} - \frac{1}{\mu_2}\right) - \left(\frac{n_1}{\lambda_2} - \frac{1}{\mu_1}\right) > 0,$$

we have

$$(1-\alpha)\left(n_2 - \frac{\lambda_1}{\mu_2}\right) - \alpha\left(n_1 - \frac{\lambda_2}{\mu_1}\right) = \alpha(1-\alpha)\lambda c = O(n),$$

while the standard deviations are $O(\sqrt{n})$. Define the middle point $M = \left((1-\alpha)\left(n_2 - \frac{\lambda_1}{\mu_2}\right) + \alpha\left(n_1 - \frac{\lambda_2}{\mu_1}\right)\right)/2$. As $n \to \infty$, we have $P(\alpha\tilde{I}_1 \geq M) = o\left(\frac{1}{\sqrt{n}}\right)$ and $P((1-\alpha)\tilde{I}_2 \leq M) = o\left(\frac{1}{\sqrt{n}}\right)$. Therefore,

$$\begin{aligned} P(\alpha I_1 \geq (1-\alpha)I_2) &\leq P(\alpha I_1 \geq M) + P((1-\alpha)I_2 \leq M) \leq P(\alpha\tilde{I}_1 \geq M) \\ &+ P((1-\alpha)\tilde{I}_2 \leq M) = o\left(\frac{1}{\sqrt{n}}\right). \end{aligned}$$

□

Proof of Theorem 8 Given $i_1 \in (0, \theta)$, $i_2 \in (0, 1-\theta)$, and $i_2 > \frac{\alpha}{1-\alpha} i_1$, for $0 \le k < i_2$,

$$\begin{aligned}
&P(K = kn | I_1 = i_1 n, I_2 = i_2 n) \\
&= B_2 \binom{n_1}{i_1 n} \binom{n_2}{i_2 n} \frac{i_1 n (i_2 n)! (i_1 n + i_2 n - kn - 1)!}{(i_2 n - kn)!} \mu_1^{i_1 n} \mu_2^{i_2 n} \lambda^{-i_1 n - i_2 n} \alpha^{-kn} \\
&\sim B_3 \left((i_1 + i_2 - k)(i_2 - k)\right)^{-1/2} \frac{\left((i_1 + i_2 - k)n\right)^{(i_1 + i_2 - k)n}}{\left((i_2 - k)n\right)^{(i_2 - k)n}} \alpha^{-kn},
\end{aligned}$$

where $B_2 = B_1 / P(I_1 = i_1 n, I_2 = i_2 n)$, $B_3 = B_2 \binom{n_1}{i_1 n} \binom{n_2}{i_2 n} i_1 (i_2 n)! \left(\frac{\mu_1}{\lambda}\right)^{i_1 n} \left(\frac{\mu_2}{\lambda}\right)^{i_2 n} \exp(-i_1 n)$. Choose k to maximize

$$\left((i_1 + i_2 - k)n\right) \log\left((i_1 + i_2 - k)n\right) - \left((i_2 - k)n\right) \log\left((i_2 - k)n\right) - kn \log \alpha.$$

The first-order condition is

$$-\log\left((i_1 + i_2 - k)n\right) + \log\left((i_2 - k)n\right) - \log \alpha = 0.$$

Therefore, the optimal value is

$$k = \bar{k} = i_2 - \frac{\alpha}{1 - \alpha} i_1.$$

Given $K = \bar{k} n + x\sqrt{n}$,

$$\begin{aligned}
&\left((i_1 + i_2)n - (\bar{k} n + x\sqrt{n})\right) \log\left((i_1 + i_2)n - (\bar{k} n + x\sqrt{n})\right) \\
&= \left((i_1 + i_2 - \bar{k})n\right) \log\left((i_1 + i_2 - \bar{k})n\right) - x\sqrt{n} \left(\log\left((i_1 + i_2 - \bar{k})n\right) + 1\right) \\
&\quad + \frac{x^2}{2(i_1 + i_2 - \bar{k})}, \left(i_2 n - (\bar{k} n + x\sqrt{n})\right) \log\left(i_2 n - (\bar{k} n + x\sqrt{n})\right) \\
&= \left((i_2 - \bar{k})n\right) \log\left((i_2 - \bar{k})n\right) - x\sqrt{n} \left(\log\left((i_2 - \bar{k})n\right) + 1\right) + \frac{x^2}{2(i_2 - \bar{k})}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\log \frac{P(K = \bar{k} n + x\sqrt{n} | I_1 = i_1 n, I_2 = i_2 n)}{P(K = \bar{k} n | I_1 = i_1 n, I_2 = i_2 n)} \sim -\frac{i_1 x^2}{2(i_1 + i_2 - \bar{k})(i_2 - \bar{k})} \\
&= -\frac{x^2}{2\alpha i_1 / (1 - \alpha)^2}.
\end{aligned}$$

Therefore, $\frac{K - \bar{k} n}{\sqrt{n}} \Big| I_1 = i_1 n, I_2 = i_2 n$ is a normal distribution with mean 0 and variance

$$\sigma_K^2 = \frac{\alpha i_1}{(1 - \alpha)^2}.$$

□

Proof of Theorem 9 From Theorem 8, we know that, as $n \to \infty$, the percentage of type c_1 customers served by type s_1 servers goes to 0 faster than $O(\frac{1}{\sqrt{n}})$. Therefore, as $n \to \infty$, the two server pools are decoupled in the sense that type s_1 servers serving type c_1 customers do not affect the fluid and diffusion limits of the decoupled systems. From the proof of Proposition 1, we know $\left(I_1 - (n_1 - \frac{\lambda_2}{\mu_1})\right)/\sqrt{n}$ converges to a normal distribution with mean 0 and variance $\frac{\lambda_2}{n\mu_1}$; independently, $\left(I_2 - (n_2 - \frac{\lambda_1}{\mu_2})\right)/\sqrt{n}$ converges to a normal distribution with mean 0 and variance $\frac{\lambda_1}{n\mu_2}$. □

E Appendix: Proofs for Sect. 4.5

Proof of Proposition 2 We want to show $\frac{P(K(\alpha_2)=k)}{P(K(\alpha_1)=k)}$ is decreasing in k. Note that

$$P(K(\alpha) = k) = \sum_{i_1=1}^{n_1} \sum_{i_2=k}^{n_2} \pi(k, i_1, i_2) + \sum_{i_2=1}^{k} \pi(k, 0, i_2) + \pi(k, 0, 0).$$

From Theorem 2, given k, for any $i_1 \in \{1, \ldots, n_1\}$, $i_2 \in \{k, \ldots, n_2\}$,

$$\frac{\pi_{\alpha_2}(k, i_1, i_2)}{\pi_{\alpha_1}(k, i_1, i_2)} = \frac{B_1(\alpha_2)}{B_1(\alpha_1)} \left(\frac{\alpha_1}{\alpha_2}\right)^k,$$

which is decreasing in k; for any $i_2 = \{1, \ldots, k\}$,

$$\frac{\pi_{\alpha_2}(k, 0, i_2)}{\pi_{\alpha_1}(k, 0, i_2)} = \frac{B_1(\alpha_2)}{B_1(\alpha_1)} \prod_{j=n-k}^{n-i_2} \frac{\mu_1 n_1 + \mu_2(j - n_1) - (1 - \alpha_1)\lambda}{\mu_1 n_1 + \mu_2(j - n_1) - (1 - \alpha_2)\lambda} \left(\frac{\alpha_1}{\alpha_2}\right)^{i_2},$$

which is decreasing in k;

$$\frac{\pi_{\alpha_2}(k, 0, 0)}{\pi_{\alpha_1}(k, 0, 0)} = \frac{B_1(\alpha_2)}{B_1(\alpha_1)} \prod_{j=n-k}^{n-1} \frac{\mu_1 n_1 + \mu_2(j - n_1) - (1 - \alpha_1)\lambda}{\mu_1 n_1 + \mu_2(j - n_1) - (1 - \alpha_2)\lambda},$$

which is decreasing in k. Therefore, $\frac{P(K(\alpha_2)=k)}{P(K(\alpha_1)=k)}$ is decreasing in k, that is,

$$\frac{P(K(\alpha_1) = k + 1)}{P(K(\alpha_2) = k + 1)} > \frac{P(K(\alpha_1) = k)}{P(K(\alpha_2) = k)},$$

meaning $K(\alpha_1)$ is larger than $K(\alpha_2)$ in the likelihood ratio order, implying $K(\alpha_1)$ stochastically dominates $K(\alpha_2)$. □

Proof of Corollary 1 The condition $\frac{n_2}{\lambda_1} - \frac{1}{\mu_2} = \frac{n_1}{\lambda_2} - \frac{1}{\mu_1}$ is equivalent to $\alpha = 1 - \beta$. By Proposition 2 $K_{\alpha_1} \geq_{ST} K_{1-\beta} \geq_{ST} K_{\alpha_2}$ whenever $\alpha_1 < 1 - \beta < \alpha_2$. But, for all $1 - \beta < \alpha_2$, $K_{\alpha_2}/n \to 0$, and for $\alpha_1 < 1 - \beta$, $\lim_{\alpha_1 \to 1-\beta} \lim_{n\to\infty} K_{\alpha_1}/n = 0$, and the corollary follows. □

Proof of Theorem 10 We prove this theorem in two steps:

- Prove that fluid limits are $\lim_{n\to\infty} I_1/n = f_{1,k}$, $\lim_{n\to\infty} I_2/n = f_{2,k}$.
- Prove the central limit behavior.

When $i_1 \in [m_1 - \epsilon n, m_1 + \epsilon n]$ and $i_2 \in [m_2 - \epsilon n, m_2 + \epsilon n]$, and n grows large, we can use Stirling's approximation:

$$
\begin{aligned}
&P(I_1 = i_1, I_2 = i_2 | K = kn) \\
&= B_{2,k} \binom{n_1}{i_1} \binom{n_2}{i_2} \frac{i_1 i_2! (i_1 + i_2 - kn - 1)!}{(i_2 - kn)!} \mu_1^{i_1} \mu_2^{i_2} \lambda^{-i_1 - i_2} \alpha^{-kn} \\
&= B_{2,k} \frac{i_1}{i_1 + i_2 - kn} \frac{n_1! n_2! (i_1 + i_2 - kn)!}{(n_1 - i_1)! i_1! (n_2 - i_2)! (i_2 - kn)!} \left(\frac{\mu_1}{\lambda}\right)^{i_1} \left(\frac{\mu_2}{\lambda}\right)^{i_2} \alpha^{-kn} \\
&\sim B_{3,k} \left(\frac{i_1}{(i_1 + i_2 - kn)(n_1 - i_1)(n_2 - i_2)(i_2 - kn)} \right)^{1/2} \\
&\quad \frac{(i_1 + i_2 - kn)^{i_1 + i_2 - kn} \exp(-i_1 - i_2)}{(n_1 - i_1)^{n_1 - i_1} i_1^{i_1} (n_2 - i_2)^{n_2 - i_2} (i_2 - kn)^{i_2 - kn}} \left(\frac{\mu_1}{\lambda}\right)^{i_1} \left(\frac{\mu_2}{\lambda}\right)^{i_2} \\
&= B_{3,k} \left(\frac{i_1}{(i_1 + i_2 - kn)(n_1 - i_1)(n_2 - i_2)(i_2 - kn)} \right)^{1/2} \\
&\quad \exp\big((i_1 + i_2 - kn) \log(i_1 + i_2 - kn) - (n_1 - i_1) \log(n_1 - i_1) \\
&\quad - i_1 \log(i_1) - (n_2 - i_2) \log(n_2 - i_2) - (i_2 - kn) \log(i_2 - kn) \\
&\quad + i_1 \log\left(\frac{\mu_1}{\lambda}\right) + i_2 \log\left(\frac{\mu_2}{\lambda}\right) - i_1 - i_2 \big) \\
&= B_{4,k} \exp\Big(n \big((x_1 + x_2 - k) \log(x_1 + x_2 - k) - (\theta - x_1) \log(\theta - x_1) \\
&\quad - x_1 \log(x_1) - (1 - \theta - x_2) \log(1 - \theta - x_2) \\
&\quad - (x_2 - k) \log(x_2 - k) + x_1 \log\left(\frac{\mu_1}{r}\right) + x_2 \log\left(\frac{\mu_2}{r}\right) - x_1 - x_2 \big) \Big),
\end{aligned}
$$

where $B_{2,k} = B_1 / P(K = kn)$, $B_{3,k} = B_{2,k} n_1! n_2! (2\pi)^{-\frac{3}{2}} \mathrm{e}^n \alpha^{-kn}$, $B_{4,k} = B_3 n^{-n-3/2} \left(\frac{x_1}{(x_1 + x_2 - k)(\theta - x_1)(1 - \theta - x_2)(x_2 - k)} \right)^{1/2}$, $x_1 = \frac{i_1}{n}$, $x_2 = \frac{i_2}{n}$. We define

$$
\begin{aligned}
F(x_1, x_2) &= (x_1 + x_2 - k) \log(x_1 + x_2 - k) - (\theta - x_1) \log(\theta - x_1) \\
&\quad - (1 - \theta - x_2) \log(1 - \theta - x_2) + x_1 (\log \mu_1 - \log r - \log x_1) \\
&\quad + x_2 (\log \mu_2 - \log r - \log x_2) - x_1 - x_2.
\end{aligned}
$$

The first-order derivatives on x_1 and x_2 are

$$
\begin{aligned}
\frac{\partial F}{\partial x_1} &= \log(x_1 + x_2 - k) + \log(\theta - x_1) - \log(x_1) - \log \frac{\mu_1}{r} = 0, \\
\frac{\partial F}{\partial x_2} &= \log(x_1 + x_2 - k) + \log(1 - \theta - x_2) - \log(x_2 - k) - \log \frac{\mu_2}{r} = 0.
\end{aligned}
$$

We can solve

$$x_1 = f_{1,k}, \qquad x_2 = f_{2,k}.$$

Look at the second-order derivatives:

$$\begin{aligned}
\frac{\partial^2 F}{\partial x_1^2} &= -\frac{1}{\theta - x_1} - \frac{1}{x_1} + \frac{1}{x_1 + x_2 - k} < 0,\\
\frac{\partial^2 F}{\partial x_2^2} &= -\frac{1}{1 - \theta - x_2} - \frac{1}{x_2 - k} + \frac{1}{x_1 + x_2 - k} < 0,\\
\frac{\partial^2 F}{\partial x_1 \partial x_2} &= \frac{1}{x_1 + x_2 - k},\\
\frac{\partial^2 F}{\partial x_1^2}\frac{\partial^2 F}{\partial x_2^2} - \left(\frac{\partial^2 F}{\partial x_1 \partial x_2}\right)^2 &= \frac{x_1^2(1 - \theta - k) + (x_2 - k)^2\theta}{x_1(x_2 - k)(x_1 + x_2 - k)(\theta - x_1)(1 - \theta - x_2)} > 0.
\end{aligned}$$

The Hessian matrix is negative definite. Therefore, $F(x_1, x_2)$ is strictly concave on $(0, \theta) \times (0, 1 - \theta)$ and reaches its unique global maximum at $(f_{1,k}, f_{2,k})$. Similar to the proof of Theorem 4, we can show that

$$\lim_{n\to\infty} \frac{I_1}{n} \to f_{1,k}, \quad \lim_{n\to\infty} \frac{I_2}{n} \to f_{2,k}.$$

To obtain the asymptotic distribution of I_1, I_2 as $n \to \infty$, we only need to consider i_1, i_2 for which $i_1/n \to f_{1,k}$ and $i_2/n \to f_{2,k}$. Similar to the proof of Theorem 5, we write $i_1 = f_{1,k}n + z_1\sqrt{n}$, $i_2 = f_{2,k}n + z_2\sqrt{n}$, with $z_1/\sqrt{n} \to 0$, $z_2/\sqrt{n} \to 0$.

$$\begin{aligned}
&P(I_1 = i_1, I_2 = i_2 | K = kn) \sim\\
&B_{3,k}\left(\frac{i_1}{(i_1 + i_2 - kn)(n_1 - i_1)(n_2 - i_2)(i_2 - kn)}\right)^{1/2}\\
&\exp\big((i_1 + i_2 - kn)\log(i_1 + i_2 - kn) - (n_1 - i_1)\log(n_1 - i_1)\\
&- i_1\log(i_1) - (n_2 - i_2)\log(n_2 - i_2) - (i_2 - kn)\log(i_2 - kn)\\
&+ i_1\log\left(\frac{\mu_1}{\lambda}\right) + i_2\log\left(\frac{\mu_2}{\lambda}\right) - i_1 - i_2\big).
\end{aligned}$$

From the definitions of $f_{1,k}$ and $f_{2,k}$ in Theorem 10,

$$\frac{(\theta - f_{1,k})\mu_1}{\lambda f_{1,k}} = \frac{(1 - \theta - f_{2,k})\mu_2}{\lambda(f_{2,k} - k)} = \frac{1}{n(f_{1,k} + f_{2,k} - k)},$$

Therefore, similar to the proof of Theorem 5, we can obtain

$$\log\left(P(I_1 = i_1, I_2 = i_2|K = kn)\right) \sim B_{5,k} + \frac{(z_1+z_2)^2}{2(f_{1,k}+f_{2,k}-k)} - \frac{z_1^2\theta}{2(\theta-f_{1,k})f_{1,k}} - \frac{z_2^2(1-\theta-k)}{2(1-\theta-f_{2,k})(f_{2,k}-k)}, \tag{15}$$

where $B_{5,k} = \log\left(B_{3,k}\left(\frac{f_{1,k}}{(f_{1,k}+f_{2,k}-k)(f_{2,k}-k)(\theta-f_{1,k})(1-\theta-f_{2,k})}\right)^{1/2}\right) - \frac{5}{2}n\log n - n(\theta\log(\theta-f_{1,k}) + (1-\theta)\log(1-\theta-f_{2,k}) + f_{1,k}+f_{2,k}) + kn(\log(f_{2,k}-k) - \log(f_{1,k}+f_{2,k}-k))$. Therefore, organizing the formula, we have

$$\left(\frac{I_1 - f_{1,k}n}{\sqrt{n}}, \frac{I_2 - f_{2,k}n}{\sqrt{n}}\middle| K = kn\right) \Rightarrow N\left(0, \begin{bmatrix}\sigma_{1,k}^2 & \rho_k\sigma_{1,k}\sigma_{2,k}\\ \rho_k\sigma_{1,k}\sigma_{2,k} & \sigma_{2,k}^2\end{bmatrix}\right),$$

where

$$\rho_k = \left(\frac{f_{1,k}(f_{2,k}-k)(\theta-f_{1,k})(1-\theta-f_{2,k})}{(f_{1,k}^2+(f_{2,k}-k)\theta)((f_{2,k}-k)^2+f_{1,k}(1-\theta-k))}\right)^{\frac{1}{2}},$$

$$\sigma_{1,k} = \left(\frac{(\theta-f_{1,k})f_{1,k}((f_{2,k}-k)^2+f_{1,k}(1-\theta-k))}{f_{1,k}^2(1-\theta-k)+(f_{2,k}-k)^2\theta}\right)^{\frac{1}{2}},$$

$$\sigma_{2,k} = \left(\frac{(1-\theta-f_{2,k})(f_{2,k}-k)(f_{1,k}^2+(f_{2,k}-k)\theta)}{f_{1,k}^2(1-\theta-k)+(f_{2,k}-k)^2\theta}\right)^{\frac{1}{2}}.$$

□

Proof of Theorem 11 The density of the highest point of the approximating binormal distribution in Theorem 10 is

$$\frac{1}{\sqrt{2\pi(1-\rho_k^2)}\sigma_{1,k}\sigma_{2,k}} = \sqrt{\frac{(f_{2,k}-k)^2\theta + f_{1,k}^2(1-\theta-k)}{2\pi f_{1,k}(f_{2,k}-k)(\theta-f_{1,k})(1-\theta-f_{2,k})(f_{1,k}+f_{2,k}-k)}}.$$

For a large n, letting $\lceil x\rceil$ be the ceiling of a real number x, and from Theorem 10,

$$P(I_1 = \lceil f_{1,k}n\rceil, I_2 = \lceil f_{2,k}n\rceil|K = kn) \sim \frac{1}{n}\sqrt{\frac{(f_{2,k}-k)^2\theta + f_{1,k}^2(1-\theta-k)}{2\pi f_{1,k}(f_{2,k}-k)(\theta-f_{1,k})(1-\theta-f_{2,k})(f_{1,k}+f_{2,k}-k)}}.$$

Combined with (15), we have

$$B_{5,k} \sim -\log n + \frac{1}{2}\log\left(\frac{(f_{2,k}-k)^2\theta + f_{1,k}^2(1-\theta-k)}{2\pi f_{1,k}(f_{2,k}-k)(\theta-f_{1,k})(1-\theta-f_{2,k})(f_{1,k}+f_{2,k}-k)}\right).$$

Recall that

$$\begin{aligned}
B_{5,k} &= \log\left(B_{3,k}\left(\frac{f_{1,k}}{(f_{1,k}+f_{2,k}-k)(f_{2,k}-k)(\theta-f_{1,k})(1-\theta-f_{2,k})}\right)^{1/2}\right) - \frac{5}{2}n\log n \\
&\quad - n(\theta\log(\theta-f_{1,k}) + (1-\theta)\log(1-\theta-f_{2,k}) + f_{1,k}+f_{2,k}) \\
&\quad + kn(\log(f_{2,k}-k) - \log(f_{1,k}+f_{2,k}-k)) \\
&= -\log(P(K=kn)) + \log\left(B_1 n_1! n_2! (2\pi n)^{-3/2}\right) \\
&\quad + \frac{1}{2}\log\left(\frac{f_{1,k}}{(f_{1,k}+f_{2,k}-k)(f_{2,k}-k)(\theta-f_{1,k})(1-\theta-f_{2,k})}\right) \\
&\quad + n - \frac{5}{2}n\log n - n(\theta\log(\theta-f_{1,k}) + (1-\theta)\log(1-\theta-f_{2,k}) + f_{1,k}+f_{2,k}) \\
&\quad + kn(\log(f_{2,k}-k) - \log(f_{1,k}+f_{2,k}-k)) - kn\log\alpha.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\log P(K=kn) &\sim B_6 + \log(f_{1,k}) - \frac{1}{2}\log\left((f_{2,k}-k)^2\theta + f_{1,k}^2(1-\theta-k)\right) \\
&\quad - n\left(\theta\log(\theta-f_{1,k}) + (1-\theta)\log(1-\theta-f_{2,k}) + f_{1,k}+f_{2,k} - k\log(f_{2,k}-k)\right. \\
&\quad \left. + k\log(f_{1,k}+f_{2,k}-k) + k\log\alpha\right),
\end{aligned}$$

where

$$B_6 = \log\left(B_1 n_1! n_2! (2\pi)^{-1}\right) + n - \frac{5}{2}n\log n - \frac{3}{2}\log n.$$

Define

$$\begin{aligned}
G(k) &= \theta\log(\theta-f_{1,k}) + (1-\theta)\log(1-\theta-f_{2,k}) + f_{1,k}+f_{2,k} \\
&\quad - k\log(f_{2,k}-k) + k\log(f_{1,k}+f_{2,k}-k) + k\log\alpha.
\end{aligned}$$

From Theorem 10, we can denote k by T,

$$k = \frac{(\mu_1-\mu_2)\theta - (1+\mu_1 T)(r-\mu_2+r\mu_2 T)}{\mu_2(1+\mu_1 T)}.$$

Note that T is nonnegative and no larger than the value in (3), denoted by $\overline{T}$. Note also that $f_{1,k} = \frac{T\theta}{T+1/\mu_1}$, $f_{2,k} = \frac{T(1-\theta-k)}{T+1/\mu_2} + k$. Algebra gives

$$\frac{\mathrm{d}G}{\mathrm{d}T} = \frac{\left(r\mu_2(1+\mu_1 T)^2 + \mu_1(\mu_1-\mu_2)\theta\right)\left(\log\left(r - \frac{\mu_1\theta}{1+\mu_1 T}\right) - \log(\alpha r)\right)}{\mu_2(1+\mu_1 T)^2}.$$

Solving $\frac{\mathrm{d}G}{\mathrm{d}T} = 0$ gives

$$T = \frac{\theta}{(1-\alpha)r} - \frac{1}{\mu_1}.$$

If

$$\frac{n_2}{\lambda_1} - \frac{1}{\mu_2} < \frac{n_1}{\lambda_2} - \frac{1}{\mu_1},$$

then

$$\overline{T} < \frac{n_1}{\lambda_2} - \frac{1}{\mu_1} = \frac{\theta}{(1-\alpha)r} - \frac{1}{\mu_1},$$

and $G(T)$ is minimized at $T^\star = \overline{T}$; otherwise,

$$\overline{T} \geq \frac{n_1}{\lambda_2} - \frac{1}{\mu_1} = \frac{\theta}{(1-\alpha)r} - \frac{1}{\mu_1},$$

and $G(T)$ is minimized at $T^\star = \frac{\theta}{(1-\alpha)r} - \frac{1}{\mu_1}$. When $G(T)$ is minimized at $T^\star = \overline{T}$, the corresponding $k^\star = 0$, we go back to the pooled system case; when $G(T)$ is minimized at $T^\star = \frac{\theta}{(1-\alpha)r} - \frac{1}{\mu_1}$, the corresponding

$$k^\star = \alpha r \left(\frac{n_2}{\lambda_1} - \frac{1}{\mu_2} - \frac{n_1}{\lambda_2} + \frac{1}{\mu_1} \right).$$

By now we have shown that, for any $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left| \frac{K}{n} - k^\star \right| > \epsilon \right) = 0.$$

Therefore,

$$\lim_{n\to\infty} P\left(\left| \frac{I_1}{n} - f_1^\star \right| > \epsilon \right) = 0, \lim_{n\to\infty} P\left(\left| \frac{I_2}{n} - f_2^\star \right| > \epsilon \right) = 0,$$
$$f_1^\star = \theta - \frac{(1-\alpha)r}{\mu_1}, f_2^\star = 1 - \theta - \frac{\alpha r}{\mu_2}.$$

This is consistent with our intuitive calculation in Sect. 3.

Suppose T changes from $T^\star = \frac{\theta}{(1-\alpha)r} - \frac{1}{\mu_1}$ to $T^\star + x/\sqrt{n}$; then $f_{1,k}$ changes $\delta f_1 = f_{1,k}^{'} \frac{x}{\sqrt{n}} + f_{1,k}^{''} \frac{x^2}{2n} + o\left(\frac{1}{n}\right)$, $f_{2,k}$ changes $\delta f_2 = f_{2,k}^{'} \frac{x}{\sqrt{n}} + f_{2,k}^{''} \frac{x^2}{2n} + o\left(\frac{1}{n}\right)$, and k changes $\delta k = k^{'} \frac{x}{\sqrt{n}} + k^{''} \frac{x^2}{2n} + o\left(\frac{1}{n}\right)$,

$$
\begin{aligned}
nG(T) &= n_1 \log(\theta - f_{1,k}) + n_2 \log(1-\theta - f_{2,k}) + f_{1,k}n + f_{2,k}n - kn\log(f_{2,k}-k) \\
&\quad + kn\log(f_{1,k}+f_{2,k}-k) + kn\log\alpha \\
&= n_1\left(\log(\theta - f_1^\star) - \frac{\delta f_1}{\theta - f_1^\star} - \frac{(\delta f_1)^2}{2(\theta - f_1^\star)^2}\right) \\
&\quad + n_2\left(\log(1-\theta - f_2^\star) - \frac{\delta f_2}{(1-\theta - f_2^\star)} - \frac{(\delta f_2)^2}{2(1-\theta - f_2^\star)^2}\right) \\
&\quad + (f_1^\star + f_2^\star)n + (\delta f_1 + \delta f_2)n + \left(k^\star + \delta k\right) n\log\alpha \\
&\quad - \left(k^\star + \delta k\right) n\left(\log(f_2^\star - k^\star) + \frac{\delta f_2 - \delta k}{f_2^\star - k^\star} - \frac{(\delta f_2 - \delta k)^2}{2(f_2^\star - k^\star)^2}\right) \\
&\quad + \left(k^\star + \delta k\right) n\left(\log(f_1^\star + f_2^\star - k^\star) + \frac{\delta f_1 + \delta f_2 - \delta k}{f_1^\star + f_2^\star - k^\star} - \frac{(\delta f_1 + \delta f_2 - \delta k)^2}{2(f_1^\star + f_2^\star - k^\star)^2}\right) + o(1).
\end{aligned}
$$

The coefficient of the $x\sqrt{n}$ term is

$$
\begin{aligned}
&\frac{-\theta f_{1,k}^{'}}{\theta - f_1^\star} - \frac{(1-\theta) f_{2,k}^{'}}{1-\theta - f_2^\star} + f_{1,k}^{'} + f_{2,k}^{'} + k^{'}\log\alpha - k^{'}\log\left(f_2^\star - k^\star\right) \\
&\quad - \frac{(f_{2,k}^{'} - k^{'})k^\star}{f_2^\star - k^\star} + k^{'}\log\left(f_1^\star + f_2^\star - k^\star\right) + \frac{\left(f_{1,k}^{'} + f_{2,k}^{'} - k^{'}\right)k^\star}{f_1^\star + f_2^\star - k^\star},
\end{aligned}
$$

which equals 0. The $O(1)$ term is

$$
\begin{aligned}
&\Bigg(- \frac{\left(f_{1,k}^{'}\right)^2 \theta}{2\left(\theta - f_1^\star\right)^2} - \frac{\theta f_{1,k}^{''}}{2\left(\theta - f_1^\star\right)} - \frac{\left(f_{2,k}^{'}\right)^2 (1-\theta)}{2\left(1-\theta - f_2^\star\right)^2} - \frac{(1-\theta) f_{2,k}^{''}}{2\left(1-\theta - f_2^\star\right)} \\
&\quad + \frac{k^{''}}{2}\left(\log\alpha - \log(f_2^\star - k^\star) + \log\left(f_1^\star + f_2^\star - k^\star\right)\right) \\
&\quad + \frac{f_{1,k}^{''} + f_{2,k}^{''}}{2} - \frac{\left(f_{2,k}^{'} - k^{'}\right)k^{'}}{f_2^\star - k^\star} + \frac{\left(f_{2,k}^{'} - k^{'}\right)^2 k^\star}{2\left(f_2^\star - k^\star\right)^2} - \frac{k^\star\left(f_{2,k}^{''} - k^{''}\right)}{2\left(f_2^\star - k^\star\right)} \\
&\quad + \frac{\left(f_{1,k}^{'} + f_{2,k}^{'} - k^{'}\right)k^{'}}{f_1^\star + f_2^\star - k^\star} - \frac{\left(f_{1,k}^{'} + f_{2,k}^{'} - k^{'}\right)^2 k^\star}{2\left(f_1^\star + f_2^\star - k^\star\right)^2} + \frac{k^\star\left(f_{1,k}^{''} + f_{2,k}^{''} - k^{''}\right)}{2\left(f_1^\star + f_2^\star - k^\star\right)}\Bigg) x^2.
\end{aligned}
$$

The above equals

$$
\frac{(1-\alpha)^2 r^2((1-\alpha)^2 r(\mu_1 - \mu_2) + \mu_1\mu_2\theta)}{2\alpha\mu_1\mu_2\theta^2} x^2.
$$

Noting that

$$k^{'}\big|_{T=T^\star} = -\frac{\lambda((1-\alpha)^2 r(\mu_1-\mu_2)+\mu_1\mu_2\theta)}{\mu_1\mu_2\theta},$$

the change from $k^\star$ to $k^\star + y\sqrt{n}$ gives

$$y = k^{'}\big|_{T=T^\star} x.$$

Therefore, we have

$$\begin{aligned}\frac{P(K=k^\star n+y\sqrt{n})}{P(K=k^\star)} &\sim \frac{(1-\alpha)^2 r^2((1-\alpha)^2 r(\mu_1-\mu_2)+\mu_1\mu_2\theta)}{2\alpha\mu_1\mu_2\theta^2}\\ &\left(-\frac{\mu_1\mu_2\theta}{\lambda((1-\alpha)^2 r(\mu_1-\mu_2)+\mu_1\mu_2\theta)}y\right)^2\\ &= \frac{y^2(1-\alpha)^2\mu_1\mu_2}{2\alpha((1-\alpha)^2 r(\mu_1-\mu_2)+\mu_1\mu_2\theta)}.\end{aligned}$$

Therefore, as $n\to\infty$, the variance of $\frac{K-k^\star n}{\sqrt{n}}$ converges to

$$\sigma_K^2 = \frac{\alpha((1-\alpha)^2 r(\mu_1-\mu_2)+\mu_1\mu_2\theta)}{(1-\alpha)^2\mu_1\mu_2} = \alpha\left(r\left(\frac{1}{\mu_2}-\frac{1}{\mu_1}\right)+\frac{\theta}{(1-\alpha)^2}\right);$$

$$\frac{K-k^\star n}{\sqrt{n}} \to \mathcal{N}\left(0,\sigma_K^2\right).$$

This is consistent with our calculation in Sect. 4.4.

If

$$\frac{n_2}{\lambda_1}-\frac{1}{\mu_2} = \frac{n_1}{\lambda_2}-\frac{1}{\mu_1},$$

$k^\star = 0$, the above calculation is valid only for $x<0$, and $\frac{K}{\sqrt{n}}$ converges to a truncated normal distribution. The density function is

$$f_{\frac{K}{\sqrt{n}}}(k) = \frac{2}{\sqrt{2\sigma_K^2\pi}}\exp\left(-\frac{k^2}{2\sigma_K^2}\right), \forall k\geq 0.$$

Note that $P(K=0)\sim\frac{2}{n\sqrt{2\sigma_K^2\pi}}\to 0$ as $n\to\infty$.

If

$$\frac{n_2}{\lambda_1}-\frac{1}{\mu_2} < \frac{n_1}{\lambda_2}-\frac{1}{\mu_1},$$

$k^\star = 0$, the above calculation is no longer valid because the coefficient of the $x\sqrt{n}$ term is nonzero. From Theorem 6 we know that K converges to a geometric distribution when $n \to \infty$. □

References

1. Adan, I.J.B.F., Weiss, G.: Exact FCFS matching rates for two infinite multi-type sequences. Oper. Res. **60**(2), 475–489 (2012)
2. Adan, I.J.B.F., Weiss, G.: A queue with skill based service under FCFS–ALIS: steady state, overloaded system, and behavior under abandonments. Stoch. Syst. **4**(1), 250–299 (2014)
3. Adan, I., Foley, R., McDonald, D.: Exact asymptotics of the stationary distribution of a Markov chain: a production model. Queueing Syst. **62**(4), 311–344 (2009)
4. Adan, I., Boon, M., Weiss, G.: A design heuristic for skill based parallel service systems. arXiv preprint arXiv:1603.01404 (2014)
5. Adan, I., Busic, A., Mairesse, J., Weiss, G.: Reversibility and further properties of FCFS infinite bipartite matching. arXiv preprint arXiv:1507.05939 (2015)
6. Armony, M., Ward, A.R.: Fair dynamic routing in large-scale heterogeneous-server systems. Oper. Res. **58**(3), 624–637 (2010)
7. Bell, S.L., Williams, R.J.: Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: asymptotic optimality of a threshold policy. Ann. Appl. Probab. **11**(3), 608–649 (2001)
8. Caldentey, R., Kaplan, E.H., Weiss, G.: FCFS infinite bipartite matching of servers and customers. Adv. Appl. Probab. **41**(3), 695–730 (2009)
9. Foss, S., Chernova, N.: On the stability of a partially accessible multi-station queue with state-dependent routing. Queueing Syst. **29**(1), 55–73 (1998)
10. Ghamami, S., Ward, A.R.: Dynamic scheduling of a two-server parallel server system with complete resource pooling and reneging in heavy traffic: asymptotic optimality of a two-threshold policy. Math. Oper. Res. **38**(4), 761–824 (2013)
11. Green, L.: A queueing system with general-use and limited-use servers. Oper. Res. **33**(1), 162–182 (1985)
12. Gurvich, I., Whitt, W.: Queue-and-idleness-ratio controls in many-server service systems. Math. Oper. Res. **34**(2), 363–396 (2009)
13. Gurvich, I., Whitt, W.: Service-level differentiation in many-server service system via queue-ratio routing. Oper. Res. **58**(2), 316–328 (2010)
14. Harchol-Balter, M., Crovella, M.E., Murta, C.D.: On choosing a task assignment policy for a distributed server system. J. Parallel Distrib. Comput. **59**(2), 204–228 (1999)
15. Harrison, J.M., Lopez, M.J.: Heavy traffic resource pooling in parallel-server systems. Queueing Syst. **33**(4), 339–368 (1999)
16. Nov, Y., Weiss, G, Zhang, H.: Fluid models of parallel service systems under FCFS. arXiv preprint arXiv:1604.04497 (2016)
17. Rubino, M., Ata, B.: Dynamic control of a make-to-order, parallel-server system with cancellations. Oper. Res. **57**(1), 94–108 (2009)
18. Shanthikumar, J.G., Yao, D.D.: Comparing ordered-entry queues with heterogeneous servers. Queueing Syst. **2**(3), 235–244 (1987)
19. Tezcan, T., Dai, J.G.: Dynamic control of N-systems with many servers: asymptotic optimality of a static priority policy in heavy traffic. Oper. Res. **58**(1), 94–110 (2010)
20. Tezcan, T.: Stability analysis of N-model systems under a static priority rule. Queueing Syst. **73**(3), 235–259 (2013)
21. Visschers, J., Adan, I.J.B.F., Weiss, G.: A product form solution to a system with multi-type customers and multi-type servers. Queueing Syst. **70**(3), 269–298 (2012)
22. Ward, A.R., Armony, M.: Blind fair routing in large-scale service systems with heterogeneous customers and servers. Oper. Res. **61**(1), 228–243 (2013)
23. Williams, R.J.: On dynamic scheduling of a parallel server system with complete resource pooling. Fields Inst. Commun. **28**, 49–71 (2000)