A NEW $1/(1-\rho)$-SCALING BOUND FOR MULTISERVER QUEUES VIA A LEAVE-ONE-OUT TECHNIQUE

BY YIGE HONG[1,a]

[1]*Computer Science Department, Carnegie Mellon University* , [a]*yigeh@cs.cmu.edu*

Bounding the queue length in a multiserver queue is a central challenge in queueing theory. Even for the classic $GI/GI/n$ queue with homogeneous servers, it is highly non-trivial to derive a simple and accurate bound for the steady-state queue length that holds across all scaling regimes. A recent breakthrough by Li and Goldberg (2025) establishes bounds that scale as $1/(1-\rho)$ for any load $\rho < 1$ and number of servers n, which is the correct scaling in many well-known scaling regimes, including classic heavy-traffic, Halfin-Whitt and Nondegenerate-Slowdown. However, their bounds entail large constant factors and a highly intricate proof, suggesting room for further improvement.

In this paper, we present a new $1/(1-\rho)$-scaling bound for the $GI/GI/n$ queue. Our bound, while restricted to the light-tailed case and the first moment of the queue length, has a more interpretable and often tighter leading constant. Our proof is relatively simple, utilizing a modified $GI/GI/n$ queue, the stationarity of a quadratic test function, and a novel leave-one-out coupling technique.

Finally, we also extend our method to $GI/GI/n$ queues with fully heterogeneous service-time distributions.

1. Introduction. The $GI/GI/n$ queue is one of the most fundamental multiserver queueing models, with a wide range of applications [see, e.g., 80]. The $GI/GI/n$ queue consists of n servers and a central queue, where external jobs arrive over time and wait in the queue for service in the first-come, first-served order. When a server frees up, it picks the first job in the queue and starts serving it until the job is completed. An illustration of the $GI/GI/n$ queue is provided in Figure 1, The name "GI" refers to "general" and "independent", indicating that the interarrival times and service times are two i.i.d. sequences, where elements of each sequence follow a general distribution.

In the theory of the $GI/GI/n$ queue, a natural goal is to find a simple formula that predicts the *queue length* (the number of jobs in the queue waiting for service). A prominent line of work focuses on the *scaling* of the queue length, aiming to find asymptotically accurate bounds or approximations when the system scales up in a certain sense. This pursuit has been remarkably successful for the single-server case, dating back to the seminal work of Kingman [47]. Kingman derives a simple upper bound for the steady-state mean queue length of the $GI/GI/1$ queue, $\mathbb{E}_\pi[Q]$:

$$\mathbb{E}_\pi[Q] \le \frac{\rho^2 \mathrm{Var}[\mu S] + \mathrm{Var}[\lambda A]}{2(1-\rho)}, \tag{1}$$

where S is the service time, A is the interarrival time, $\mu \triangleq 1/\mathbb{E}[S]$, $\lambda \triangleq 1/\mathbb{E}[A]$, and $\rho \triangleq \lambda/\mu$ is the *load*. Despite its simplicity, Kingman's bound is surprisingly accurate. In particular, it is asymptotically tight in the so-called *heavy-traffic* scaling regime where $\rho \uparrow 1$ [48], so it precisely captures the queue's behavior in the scenario where the queueing delay becomes significant.

MSC2020 subject classifications: Primary 60K25; secondary 68M20, 90B22.

Keywords and phrases: multiserver queue, steady-state analysis, basic adjoint relationship, asymptotic independence, coupling.

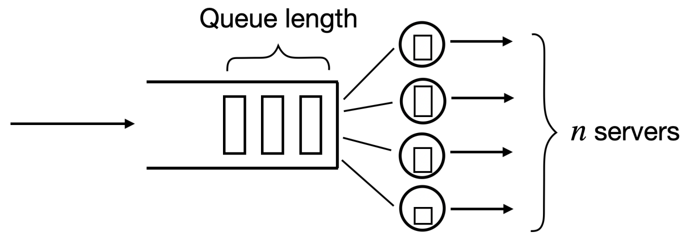

Fig 1: An illustration of $GI/GI/n$ queue, where each cycle denotes a server, and each rectangle denotes a job.

In contrast to the single-server case, finding a similarly simple and accurate upper bound for the $GI/GI/n$ queue is a much greater challenge. This complexity arises from a rich set of combinations of two basic scalings: the *heavy-traffic* scaling, where the load $\rho \uparrow 1$, and the *many-server* scaling, where the number of servers $n \to \infty$. In practice, these two parameters often scale jointly to balance utilization and delay [3, 6, 30, 57, 77], with different scaling rates giving rise to dramatically different asymptotic behaviors. For example, in the classic heavy-traffic scaling regime ($\rho \uparrow 1$ with fixed n), the limiting queue length distribution of $GI/GI/n$ queue is similar to that of $GI/GI/1$, depending only on the first two moments the input distributions [see, e.g., 50]. In contrast, in the Halfin-Whitt (HW) regime ($n \to \infty$ and $\rho = 1 - \Theta(1/\sqrt{n})$) or the Non-Degenerate Slowdown (NDS) regime ($n \to \infty$ and $\rho = 1 - \Theta(1/n)$), the limiting behavior is far more complex and can depend intricately on the full input distributions [see, e.g., 1, 4, 9, 68].

A long-standing challenge has been to establish a queue length bound with the $1/(1-\rho)$ scaling that holds *universally* for all $\rho < 1$ and n. Achieving this is significant because this scaling is known to be tight across a broad class of important operational regimes, including classic heavy-traffic, NDS, and Halfin-Whitt [1, 4, 50], although better scaling could arise when $\rho = 1 - \omega(1/\sqrt{n})$ [67]. However, most prior work focuses on particular scaling regimes without providing general upper bounds for the regimes between them (for instance, between the HW and NDS regimes), so such a universal bound remained elusive until the recent breakthrough by Li and Goldberg [53]. Their work provides the first such result, stated in their Corollary 2 as:

$$\begin{aligned} &\mathbb{E}_\pi[Q] \\ &\le \left(2.1\times 10^{21}\mathbb{E}\left[(\mu S)^2\right]\left(\mathbb{E}\left[(\mu S)^2\right]^{1+\epsilon} + \mathbb{E}\left[(\mu S)^{2+\epsilon}\right]\right)\left(\frac{1}{\epsilon}\right)^4 + 49\mathbb{E}\left[(\Lambda A)^2\right]\right)\frac{1}{1-\rho}, \end{aligned} \tag{2}$$

where $\Lambda = 1/\mathbb{E}[A]$ and $\epsilon \in (0, 1/2)$. In addition to this bound on the mean queue length, Li and Goldberg [53] also establishes several high-order moment and tail bounds with the same universal scaling.

Although Li and Goldberg [53] shows the existence of a universal bound with the correct $1/(1-\rho)$ scaling, the numerical constant in their bound is astronomically large. This stands in stark contrast to Kingman's bound for the single-server case, suggesting room for improvement. Indeed, Li and Goldberg [53] themselves conjecture that Kingman's bound may hold for any $M/GI/n$ queue. This leads to the central question of our paper: *Can we prove a universal $1/(1-\rho)$-scaling bound for the $GI/GI/n$ queue with a small constant factor?*

Our results. We prove that, in the $GI/GI/n$ queue, the steady-state mean queue length satisfies

$$\mathbb{E}_\pi[Q] \le \frac{2R_s^{\max} + \rho \mathrm{Var}[\Lambda A] - \rho}{2(1-\rho)} + \left(\frac{1}{2}\mathbb{E}\left[(\Lambda A)^2\right] - R_a^{\min}\right), \tag{3}$$

where $R_s^{\max} \triangleq \sup_{t\geq 0} \mathbb{E}[\mu S - t \mid \mu S \geq t]$ (finite for a large class of light-tailed distributions), and the second term $\mathbb{E}\left[(\Lambda A)^2\right]/2 - R_a^{\min}$ is a constant determined by the unitized arrival-time distribution. This bound holds assuming $\mathbb{E}\left[A^2\right] < \infty$, $R_s^{\max} < \infty$ (Assumption 1), S is non-lattice (Assumption 2), and a certain positive Harris recurrence condition (Assumption 3). In the special case of the $M/GI/n$ queue, (3) further simplifies into

$$\mathbb{E}_\pi[Q] \leq \frac{R_s^{\max}}{1-\rho}. \tag{4}$$

The above two bounds are given in Section 4 as direct corollaries of a more refined bound (Theorem 1), which has a better leading constant when $n = o(1/(1-\rho))$.

Our bounds scale as $1/(1-\rho)$ universally for all $\rho < 1$ and n. Compared with the bounds of [53], our bounds require stronger assumptions on the service-time distribution, but often have much smaller constants when the assumptions are satisfied. For example, in the $M/GI/n$ queue whose service times follow the gamma distribution with the shape parameter α (i.e., with the squared coefficient of variation $1/\alpha$), $R_s^{\max} = \max(1, 1/\alpha)$, so (4) implies $\mathbb{E}_\pi[Q] \leq \max(1, 1/\alpha)/(1-\rho)$. We will provide more examples in Section 4.

Our results also generalize to the $GI/GI/n$ queue with *fully heterogeneous* servers. The details can be found in Section 8.

Technical contribution. Our proof is built around a simple core argument that synthesizes several techniques. We first define a modified $GI/GI/n$ queue whose queue length stochastically dominates that of the original system, an approach also used by [53]. Our analysis of this modified queue then proceeds in two steps. First, we use the stationarity of a quadratic function to derive an expression for the mean queue length, representing it as the sum of a $GI/GI/1$ mean queue length and several "covariance" terms. Second, we bound these covariance terms.

The first step is a relatively standard application of the Basic Adjoint Relationship (BAR) [see, e.g., 11, 12, 60, 62]. Our key innovation lies in the second step, where we introduce a novel *leave-one-out coupling* to control the covariance terms. Similar covariance terms also appear in previous papers based on BAR or equivalent approaches and are one of the main obstacles to obtaining a universal bound with $1/(1-\rho)$-scaling [31, 36, 38, 72]. We provide a more detailed overview of our technical approach in Section 5.

Paper organization. We first review the related work in Section 2. In Section 3, we formally set up the model, introduce the modified $GI/GI/n$ queue, and state the assumptions about these systems. In Section 4, we state our main results and provide some examples. In Section 5, we give a high-level overview of our proof and provide intuition. In Section 6, we formally set up the analysis framework and state the supporting lemmas for our proof. In Section 7, we prove our main results for the $GI/GI/n$ queue. In Section 8, we generalize our results to the $GI/GI/n$ queue with heterogeneous servers. Finally, we conclude the paper in Section 9.

General notation. For any two real numbers a, b, we use $a \wedge b$ to denote their minimum. For any positive integer k, we let $[k] \triangleq \{1, 2, \ldots, k\}$. We let $\mathbb{R}$ be the set of real numbers, let $\mathbb{R}_{\geq 0}$ be the set of non-negative real numbers, let $\mathbb{Z}_{\geq 0}$ be the set of non-negative integers, and let $\mathbb{Z}_{\geq 1}$ be the set of positive integers. Unless otherwise specified, we use Bachmann–Landau notation with respect to the limits where $n \to \infty$, $\rho \uparrow 1$, or both simultaneously; we treat other parameters as constants and omit their dependence. We denote probability, expectation, and variance by $\mathbb{P}[\cdot]$, $\mathbb{E}[\cdot]$, and $\mathrm{Var}[\cdot]$, respectively. The subscripts of these three symbols specify the underlying probability distributions: for example, $\mathbb{P}_\nu[\cdot]$ and $\mathbb{E}_\nu[\cdot]$ are taken under the distribution ν. When the distribution is clear from context, we omit the subscript.

2. Related work.

General prior work on $GI/GI/n$ queue. There has been a rich body of literature on the queue-length scaling of $GI/GI/n$ queue. Prior to Li and Goldberg [53], most papers focus on a certain representative or useful scaling regime of ρ and n, without providing general bounds for the regimes in between. In the following, we briefly review these papers and refer the readers to [53] for a comprehensive review.

The most well-studied regime is the classic heavy-traffic regime, where n is fixed, and $\rho \uparrow 1$ [e.g., 5, 31, 36, 39, 46, 49–51, 55, 63–65, 69, 71, 74, 79]. In this regime, relatively simple bounds or approximations are available, and the queue length is known to scale as $1/(1-\rho)$ for most well-behaved instances of $GI/GI/n$ queue; exceptions arise when, for instance, service times have infinite variances [69–71, 74]. Some of these papers contain bounds that hold beyond the classic heavy-traffic regime, but they scale worse than $1/(1-\rho)$ when n is large [e.g., 31, 36, 51, 55]. For example, the bound proved by [31] scales as $O(1/(1-\rho)+n)$.

More recent papers also study the Halfin-Whitt regime (HW), where $n \to \infty$ and $\rho = 1 - \Theta(1/\sqrt{n})$ [e.g., 1, 7–10, 13, 18, 21, 27–29, 34, 35, 41–44, 56, 58, 66, 68], or the Non-Degenerate Slowdown (NDS) regime, where $n \to \infty$ and $\rho = 1 - \Theta(1/n)$ [e.g., 3, 4, 34]. Asymptotic analysis has shown that the queue length in these regimes also scales as $1/(1-\rho)$. However, there is a key distinction between the results in classic heavy-traffic results and these regimes: the analyses for the HW and NDS regimes are typically asymptotic, focusing on the system's limiting behavior as $n \to \infty$. Although some studies provide convergence rates to this limit [e.g. 9], and some prove the tightness of the $(1-\rho)$-scaled queue lengths [e.g. 18, 27–29], the error terms or bounds are generally *inexplicit* outside of certain special cases (e.g. exponential service times [8, 10, 29, 67], deterministic service times [41, 43], lattice service times [28]). It is thus difficult to translate these results into the kind of explicit, non-asymptotic performance bounds that are the focus of our work.

Some papers also consider scaling regimes with load further lighter than HW, i.e. $\rho = 1 - \omega(1/\sqrt{n})$ [e.g. 8, 10, 34, 38, 67, 75]. Such regimes are referred to as sub-HW regimes. In the sub-HW regimes, the mean queue length of the $GI/GI/n$ queue could be exponentially smaller than $1/(1-\rho)$, although the precise scaling of the mean queue length is only known for very special models, such as the $M/M/n$ queue [see, e.g. 8, 67].

Finally, apart from [53], some papers also provide results across scaling regimes, but under highly specialized distributional assumptions. Some of them assume exponential service times and provide universal approximations for all scaling regimes [e.g., 8, 10, 34, 67]. Another paper [38] provides a $1/(1-\rho)$-scaling bound for some regimes between HW and NDS, assuming Poisson arrivals and hyperexponential service times.

Basic Adjoint Relationship and similar methods. Basic Adjoint Relationship (BAR), also known as the *stationary equation* in the earlier literature [61], is a set of equations that characterize the stationary distribution of Markov processes. BAR allows us to extract information about the stationary distribution of a Markov process by taking suitable test functions. BAR and BAR-based analytical techniques have been studied in the literature under different names, such as the drift method [see, e.g., 26] or the rate conservation law [60, 62, see, e.g.,]. In this paper, we use BAR to refer to the particular framework in [11, 12], which handles Markov processes with both continuous and discrete state changes, utilizing the Palm distribution. This BAR framework has recently been applied to various queueing systems with general interarrival-time and service-time distributions under heavy traffic [e.g., 11, 14, 15, 19, 20, 22, 32, 33, 37, 62].

Some prior work has derived queue length bounds for the $GI/GI/n$ queue based on BAR or equivalent approaches [31, 36, 38, 72]. These bounds typically have small constant factors in front of the $1/(1-\rho)$-scaling term, but they also contain additive error terms that depend on n. These additive error terms cause these bounds to scale worse than $1/(1-\rho)$ when n is large, reflecting a common technical challenge in analyzing the $GI/GI/n$ queue using BAR-based approaches. We will discuss this challenge in more detail in Section 5.

Leave-one-out techniques. Our core technical innovation, the leave-one-out technique, is similar in spirit to the leave-one-out cross-validation in statistics [52], as well as some analytical techniques in learning theory [73] and matrix completion [24] that bear the same name. On a high level, these methods allow us to isolate the effect of an input component on a stochastic procedure by removing or replacing this component, and compare the changes in the outcome.

3. Problem setup. In this section, we define two queueing systems with homogeneous servers: the $GI/GI/n$ queue, whose queue length we want to upper bound, and a proxy system named *modified $GI/GI/n$ queue*, which will be the focus of our analysis.

3.1. *$GI/GI/n$ queue.* We consider the multiserver queue with general independent interarrival-time and service-time distributions, also known as the $GI/GI/n$ queue, under the first-come, first-served (FCFS) scheduling policy. As illustrated in Figure 1, this queueing system consists of a central queue and n identical servers, serving externally arriving jobs with independent and identically distributed (i.i.d.) interarrival times. Each job joins the end of the queue if all servers are busy, or starts service immediately if there are any idle servers. After entering the service, the job picks an idle server uniformly at random, occupies the server for a random amount of service time, and leaves the system after completing the service. The service times of the jobs are also i.i.d. When a server completes a job, it either immediately takes the next job from the head of the queue to serve it if the queue is non-empty, or becomes idle if the queue is empty. If multiple servers complete the service simultaneously, we assign the head-of-line jobs to these servers uniformly at random.

The $GI/GI/n$ system starts running at the time $t = 0$, possibly with some initial jobs in the queue and in service. We allow arbitrary initial queue length (denoted as $Q(0)$), time until first arrival, the set of initially busy servers, and the residual service times of the busy servers. These data describe the status of the system at the time $t = 0$, and we refer to them collectively as the *initial state*.

The $GI/GI/n$ queue is driven by two independent i.i.d. sequences of random variables, which represent the interarrival times and service times. Specifically, under the FCFS policy, we can index all jobs by the order in which they *begin service*. For the k-th job that begins service after time 0, we let S^k denote its service time and let A^k denote the interarrival time between the k-th and $(k+1)$-th jobs. Note that $\{A^k\}_{k=1}^{Q(0)}$ correspond to jobs already in the queue at time 0, so they do not play a role in the analysis and can take arbitrary values. By the definition of the $GI/GI/n$ queue, the interarrival times $\{A^k\}_{k=Q(0)+1}^{\infty}$ and the service times $\{S^k\}_{k=1}^{\infty}$ are independent, and each sequence is i.i.d. with its own distribution. We represent the distributions of A^k and S^k using two generic random variables A and S, respectively, and simply refer to them as the *interarrival time* and the *service time*. The *arrival rate* Λ and *service rate* μ of the system can therefore be defined as $\Lambda \triangleq 1/\mathbb{E}[A]$ and $\mu \triangleq 1/\mathbb{E}[S]$. We also defined the *scaled arrival rate* λ to be $\lambda \triangleq \Lambda/n$, which represents the average arrival rate processed by each server. To characterize the tails of the unitized service time μS and the unitized interarrival time ΛA, we consider the supremum of μS's mean residual time $R_s^{\max} \triangleq \sup_{t\geq 0} \mathbb{E}[\mu S - t \mid \mu S \geq t]$, and the infimum of ΛA's mean residual time $R_a^{\min} = \inf_{t\geq 0} \mathbb{E}[\Lambda A - t \mid \Lambda A \geq t]$.

Next, we state two assumptions on the distributions of A and S.

ASSUMPTION 1. We assume that $\mathbb{E}\left[A^2\right] < \infty$ and $R_s^{\max} < \infty$.

ASSUMPTION 2. We assume that the distribution of S is non-lattice, i.e., the support of S is not a subset of $\{\delta, 2\delta, 3\delta, \dots\}$ for any $\delta > 0$.

Assumption 1 requires that the interarrival time has a finite second moment, and the service time S has a bounded mean residual time. The assumption $R_s^{\max} < \infty$ has also been used in [25, 31]. As shown in [25], $R_s^{\max} < \infty$ implies that the tail CDF of S decays at least exponentially and hence that all moments of S are finite; for completeness, we also provide a proof in Section A. The mean residual time is bounded for many common classes of light-tailed distributions, including the phase-type, Increasing Failure Rate (IFR), New Better Than Used in Expectation (NBUE), gamma distributions, etc. We will show $R_s^{\max}$ of some example distributions in Section 4.

Assumption 2 can be roughly viewed as a continuous-time analogue of the aperiodicity condition for renewal processes with the inter-event distribution S. The non-lattice condition is required for some common versions of renewal theorem [2, Theorem V.4.3], which we will use in the proof of Lemma 4.

The $GI/GI/n$ queue is a Markov process whose *state* at time t can be represented as

$$X(t) \triangleq (R_{s,1}(t), R_{s,2}(t), \ldots R_{s,n}(t), R_a(t), Q(t)), \tag{5}$$

where $Q(t)$ denotes the *queue length* (not including the jobs in service), $R_a(t)$ denotes the *residual arrival time*, and $R_{s,i}(t)$ denotes the *residual service time* of the i-th server, for $i = 1, 2, \ldots, n$. We use the convention that all the state variables have right-continuous sample paths, and let $R_{s,i}(t) = 0$ if the i-th server is empty at time t. The set of all possible realizations of $X(t)$ constitutes the *state space* of $GI/GI/n$ queue, denoted as $\mathcal{X} \triangleq \mathbb{R}_{\geq 0}^{n+1} \times \mathbb{Z}_{\geq 0}$.

The load of the $GI/GI/n$ system is defined as $\rho \triangleq \Lambda/(n\mu) = \lambda/\mu$. We assume that $\rho < 1$, which implies that the system is positive Harris recurrent [2, Theorem XII.2.2].

3.2. *Modified $GI/GI/n$ queue.* Now we set up the *modified $GI/GI/n$* queue, which we will focus on in our analysis as a proxy of the original $GI/GI/n$ queue. On a high-level, the modified $GI/GI/n$ queue adds virtual jobs to the $GI/GI/n$ queue to keep the servers busy. Specifically, when the queue is empty and a server is about to idle, the server immediately begins to serve a virtual job with a service time independently sampled from the distribution of S. The server then works non-preemptively on this virtual job, becoming available to serve the next job only after completing the virtual job.

The modified $GI/GI/n$ queue has been studied by several papers ([5, 16, 35, 40, 45]; [see also 78, Chapter 10]), and it is known that given the same initial states, the queue length distribution of the modified $GI/GI/n$ stochastically dominates that of the original $GI/GI/n$ at any time. This fact has been proved in, for example, Proposition 1 of [27], and we will also include a proof in Section C for completeness. Due to this relation, an upper bound for the modified $GI/GI/n$ queue automatically holds for the original $GI/GI/n$ queue, so it suffices to focus on the modified $GI/GI/n$ queue. This approach has been one of the key steps in [53], and we will adopt the same approach in our analysis.

A nice property of the modified $GI/GI/n$ queue is that its sample path is *determined* by $(n+1)$ mutually independent renewal processes [27, Corollary 1]: the service-completion of each server is a renewal process with the inter-event distribution S, and the arrival process is a renewal process with the inter-event distribution A.

The modified $GI/GI/n$ queue is a Markov process whose state can be represented by the same tuple $X(t)$ as the original $GI/GI/n$ queue (5), with the state space $\mathcal{X} = \mathbb{R}_{\geq 0}^{n+1} \times \mathbb{Z}_{\geq 0}$. The dynamics of this Markov process is summarized below:

- **Continuous dynamics.** At any time t, the residual times (i.e., forward recurrence times) of the renewal processes, $R_a(t)$ and $R_{s,i}(t)$ for $i \in [n]$, decrease at the constant rate 1.
- **Discrete events.** When one or more residual times reach zero at time t, discrete events are triggered. Each type of event corresponds to a subroutine that maps an input state, $X^- = (R_{s,1}^-, \ldots R_{s,n}^-, R_a^-, Q^-)$, to an output state, $X^+ = (R_{s,1}^+, \ldots R_{s,n}^+, R_a^+, Q^+)$, according to the following rules:

(i) **Arrival event** ($R_a(t-)=0$): $R_a^+ = R_a^- + A$ for some interarrival time A independent of X^-; $Q^+ = Q^- + 1$; other coordinates of the state remain unchanged.

(ii) **Completion event at server** i ($R_{s,i}(t-)=0$): $R_{s,i}^+ = R_{s,i}^- + S$ for some service time S independent of X^-; $Q^+ = Q^- - \mathbb{1}\{Q^- > 0\}$; other coordinates of the state remain unchanged.

Typically, only one event happens at a time, and the post-event state $X(t)$ is generated by applying the corresponding subroutine to the pre-event state $X(t-)$. In rare cases where multiple events occur simultaneously, the subroutines are applied sequentially. Different orders of applying the subroutines result in different system dynamics, so we stick to a fixed order to avoid ambiguities: we first apply the arrival subroutine if there is an arrival event, and then any completion subroutines in increasing order of their server indices.

Next, we make a stability assumption for the modified $GI/GI/n$ queues. Note that in addition to the main modified $GI/GI/n$ queue under study, we will also construct a few more auxiliary instances of modified $GI/GI/n$ queues (Section 6.1); these auxiliary systems have different numbers of servers but the same interarrival-time and service-time distributions after proper scalings. Our assumption should also hold for all these systems; the load of each of these systems is defined as the ratio of its arrival rate and the total service rate of its servers.

ASSUMPTION 3. Each modified $GI/GI/n$ queue with homogeneous servers considered in this paper is positive Harris recurrent if its load is less than 1. In addition, let ν be its unique stationary distribution, then $\lim_{T\to\infty} \frac{1}{T}\int_0^T \mathbb{E}\left[Q(t)\right] dt = \mathbb{E}_\nu[Q] < \infty$.

REMARK 1. Assumption 3 has two parts: the positive Harris recurrence of the modified $GI/GI/n$ queue, and the finiteness of its steady-state mean queue length. Essentially, only the first part is restrictive, as the finiteness of the steady-state mean queue length could be implied by the existing queue length bounds for modified $GI/GI/n$ queues. In particular, Corollary 2 in [53] holds for any modified $GI/GI/n$ queue with load less than 1 that satisfies positive Harris recurrent, aperiodicity, and some standard finite moment assumptions implied by Assumption 1. We will discuss Assumption 3 in more detail in Section B, where we argue that certain regularity conditions on the distributions of A and S are likely to imply Assumption 3.

4. Main results.

In this section, we state our main results for the $GI/GI/n$ queue with homogeneous servers under Assumptions 1, 2 and 3, and provide some examples. We will also compare our results with those of [53]. We will present generalizations to the heterogeneous setting in Section 8.

We begin with our main bound, presented in Theorem 1 below. This bound is a more refined version of the bounds (3) and (4) from the introduction.

THEOREM 1. *Consider the $GI/GI/n$ queue with homogeneous servers under the FCFS policy, whose service-time and interarrival-time distributions satisfy Assumptions 1, 2 and 3. When the load $\rho < 1$, let π be the unique stationary distribution of the system. Then the steady-state mean queue length, $\mathbb{E}_\pi[Q]$, is bounded as*

$$\begin{aligned}\mathbb{E}_\pi[Q] \le{}& \frac{\rho \mathrm{Var}\left[\Lambda A\right] + \mathrm{Var}\left[\mu S\right] + 1 - \rho}{2(1-\rho)} + \min\left(n, \frac{1}{1-\rho}\right)\left(R_s^{\max} - \frac{1}{2}\mathbb{E}\left[(\mu S)^2\right]\right) \\ &+ \left(\frac{1}{2}\mathbb{E}\left[(\Lambda A)^2\right] - R_a^{\min}\right).\end{aligned} \tag{6}$$

We first comment on the scaling of the bound in (6). Observe that this bound only depends on the distributions of the *unitized* service time μS and the unitized interarrival time ΛA. Consequently, when we fix the distributions of μS and ΛA, this bound scales as $1/(1-\rho)$ *universally for all* $\rho < 1$ *and* n. When specialized to the Halfin-Whitt regime [35] and the Non-Degenerate Slowdown regime [3], the bound scales $O(\sqrt{n})$ and $O(n)$, respectively, which match the correct scalings in these two regimes. The "min" in the second term of (6) leads to a smaller leading constant when $n = o\big(1/(1-\rho)\big)$, which includes the classic heavy-traffic regime with n fixed.

As we might notice, the bound in (6) shares some similarities to Kingman's approximation for the $GI/GI/1$ queue [48], so it is natural wonder what each term in (6) represents and where it comes from. In the following, we provide some high-level comments and leave more detailed discussions to the technical overview (Section 5).

1. The first term in (6) matches Kingman's approximation for the $GI/GI/1$ queue whose server with interarrival time A and service time S/n, up to low-order errors; this $GI/GI/1$ queue can be interpreted as having a "super-server" that is n-times faster than each server in the $GI/GI/n$ queue. Because the first term in (6) is dominant as $\rho \uparrow 1$ with n fixed, it reflects the shared asymptotical behavior of $GI/GI/n$ and $GI/GI/1$ in the classic heavy-traffic regime. This connection between $GI/GI/n$ and $GI/GI/1$ is well known in the literature [see, e.g., 31, 36, 72]. In our analysis, this comparison between $GI/GI/n$ and $GI/GI/1$ is performed *implicitly* using the BAR technique.
2. The second and third terms in (6) bound the difference between the (modified) $GI/GI/n$ queue and the $GI/GI/1$ queue. In particular, the second term in (6) is associated with the conditional distribution of the residual service time given $Q = 0$ in the modified $GI/GI/n$ queue, reflecting a certain "boundary effect" when the queue is empty. Note that this term vanishes when S is exponentially distributed, since $\mathbb{E}\left[(\mu S)^2\right]/2$ is the expectation of μS's excess distribution, and recall that $R_s^{\max} = \sup_{t\geq 0}\mathbb{E}\left[\mu S - t \mid \mu S \geq t\right]$.
3. The third term in (6) is associated with the conditional distribution of the residual interarrival time given $Q = 0$ in the modified $GI/GI/n$ queue, and it vanishes when A is exponentially distributed.

Theorem 1 implies the following simpler bound after relaxing the minimum in (6) to $1/(1-\rho)$:

$$\mathbb{E}_\pi[Q] \leq \frac{\rho \mathrm{Var}\left[\Lambda A\right] - \rho + 2R_s^{\max}}{2(1-\rho)} + \frac{1}{2}\mathbb{E}\left[(\Lambda A)^2\right] - R_a^{\min}. \tag{7}$$

When we consider the $M/GI/n$ queue, the bound in (7) can be further simplified into:

$$\mathbb{E}_\pi[Q] \leq \frac{R_s^{\max}}{1-\rho}. \tag{8}$$

Next, we apply the bound in (8) to some examples. We assume that the examples under consideration satisfy Assumption 3.

EXAMPLE 1 (NBUE distribution). Consider the $M/GI/n$ queue whose service-time distribution is non-lattice and New Better Than Used in Expectation (NBUE). NBUE is a large class of distribution that contains all Increasing Failure Rate distributions; a distribution is NBUE if its mean residual time conditioned on any age is no larger than its original mean. Then by definition, we have $R_s^{\max} \leq \mathbb{E}\left[\mu S\right] = 1$, so (8) implies $\mathbb{E}_\pi[Q] \leq 1/(1-\rho)$.

EXAMPLE 2 (Gamma distribution). Consider the $M/GI/n$ queue with gamma-distributed service times. Suppose S follows the gamma distribution with shape parameter α and scale parameter θ, then the unitized service time μS is also gamma-distributed with shape α and scale $1/\alpha$. When $\alpha > 1$, μS has increasing failure rates, so $R_s^{\max} = \mathbb{E}[\mu S] = 1$. When $\alpha < 1$, direct calculation shows that $R_s^{\max} = 1/\alpha$. Because the gamma distribution is always non-lattice, (8) implies $\mathbb{E}_\pi[Q] \le \max(1/\alpha, 1)/(1-\rho)$.

EXAMPLE 3 (Phase-type distribution). Consider the $M/GI/n$ queue whose service-time distribution follows the phase-type distribution with d phases. For $k \in [d]$, let τ_i be the expected residual time of S conditioned on it being in the phase k. Then it is not hard to see that $R_s^{\max} \le \mu \max_{k\in[d]} \tau_k < \infty$. Moreover, phase-type distributions are non-lattice. Therefore, (8) implies $\mathbb{E}_\pi[Q] \le \mu \max_{k\in[d]} \tau_k/(1-\rho)$.

EXAMPLE 4 (Bounded distribution). Consider the $M/GI/n$ queue whose service-time distribution is non-lattice and bounded by some constant $M > 0$. Then we have $R_s^{\max} \le \mu M$, so (8) implies $\mathbb{E}_\pi[Q] \le \mu M/(1-\rho)$.

Comparing with Li and Goldberg [53]. Li and Goldberg [53] has proved a wide range of queue length bounds that scale universally as $1/(1-\rho)$. Those bounds are primarily in the form of tail bounds (Theorems 1,2, and 3), which imply bounds on the steady-state mean queue length $\mathbb{E}_\pi[Q]$ as corollaries. Here, we restate one of their bounds on $\mathbb{E}_\pi[Q]$ as an example:

THEOREM (Adapted from Corollary 2 of [53]). *Consider the $GI/GI/n$ queue with homogeneous servers under the FCFS policy. Assume that (1) $\mathbb{E}\left[A^2\right] < \infty$; (2) there exists $\epsilon \in (0, 0.5)$ s.t. $\mathbb{E}\left[S^{2+\epsilon}\right] < \infty$; (3) $\rho < 1$; (4) the system has a unique limiting distribution π. Then*

$$\begin{aligned} &\mathbb{E}_\pi[Q] \\ &\le \left(2.1\times 10^{21}\mathbb{E}\left[(\mu S)^2\right]\left(\mathbb{E}\left[(\mu S)^2\right]^{1+\epsilon} + \mathbb{E}\left[(\mu S)^{2+\epsilon}\right]\right)\left(\frac{1}{\epsilon}\right)^4 + 49\mathbb{E}\left[(\Lambda A)^2\right]\right)\frac{1}{1-\rho}, \end{aligned} \tag{9}$$

The bound in (9) achieves the $1/(1-\rho)$-scaling with the constant factor only depending on $\mathbb{E}\left[(\Lambda A)^2\right]$, $\mathbb{E}\left[(\mu S)^2\right]$ and $\mathbb{E}\left[(\mu S)^{2+\epsilon}\right]$. However, this bound involves a huge numerical constant. Similar large numerical constants also appear in all other bounds in [53] that achieve the $1/(1-\rho)$-scaling. Our bounds improve upon the bounds in [53] when $R_s^{\max}$ is reasonably small.

Finally, note that in addition to the assumption that $R_s^{\max} < \infty$, our non-lattice assumption (Assumption 2) is also not needed in [53]. This technical condition ensures that the residual service times converge to the equilibrium distribution in a proper sense, which is needed in our proof of Lemma 4. [53] does not need this assumption partly due to their model setup, where the initial residual service times are independently sampled from the equilibrium distribution.

5. Proof overview and intuition. In this section, we overview the main ideas in the proof of Theorem 1. To make the overview accessible, we start from the special cases, the $M/M/1$ queue and the $GI/GI/1$ queue, and gradually build towards the $GI/GI/n$ queue. Each step of the generalizations corresponds to a refinement of the proof techniques, and the novelty of this paper lies in the last step where we generalize from $GI/GI/1$ to $GI/GI/n$.

M/M/1 analysis: quadratic test function. Consider the M/M/1 queue with the arrival rate λ and service rate μ. Although the stationary distribution of M/M/1 is straightforward to calculate, here we present an alternative way of bounding its steady-state mean queue length $\mathbb{E}_\pi[Q]$. Recall that the queue length of the M/M/1 queue, Q, jumps to $Q + \Delta Q$ at the rate $\lambda + \mu$, where the size of the jump ΔQ conditioned on Q has the distribution

$$\Delta Q = \begin{cases} 1 & \text{with probability } \lambda/(\lambda+\mu) \\ -\mathbb{1}\{Q > 0\} & \text{with probability } \mu/(\lambda+\mu). \end{cases}$$

Now consider the quadratic function Q^2: The instantaneous expected rate of change (so-called "drift") of Q^2 is given by

$$\begin{aligned}(\lambda+\mu)\mathbb{E}\left[(Q+\Delta Q)^2 - Q^2 \,\middle|\, Q\right] &= 2Q\,(\lambda+\mu)\mathbb{E}\left[\Delta Q \,\middle|\, Q\right] + (\lambda+\mu)\mathbb{E}\left[\Delta Q^2 \,\middle|\, Q\right] \\ &= 2Q\left(\lambda - \mu\mathbb{1}\{Q>0\}\right) + \lambda + \mu\mathbb{1}\{Q>0\} \\ &= -2(\mu-\lambda)Q + \lambda + \mu\mathbb{1}\{Q>0\}.\end{aligned}$$

Observe that when Q follows its stationary distribution π, Q^2 is stationary, so the drift of Q^2 is zero, i.e.,

$$2(\mu-\lambda)\mathbb{E}_\pi[Q] = \lambda + \mu\mathbb{P}_\pi\left[Q>0\right]. \tag{10}$$

Because $\mathbb{P}_\pi[Q>0] = \rho$, we get $\mathbb{E}_\pi[Q] = \lambda/(\mu-\lambda) = \rho/(1-\rho)$.

The rationale behind using the quadratic test function Q^2 is that the leading term of its drift is almost a constant proportion of Q, because $\mathbb{E}[\Delta Q \mid Q] = \left(\lambda - \mu\mathbb{1}\{Q>0\}\right)/(\lambda+\mu)$ is almost constant except when $Q = 0$. Intuitively, the queue length drifts toward zero at a constant rate whenever the queue is busy. Consequently, considering the drift of Q^2 yields a formula for $\mathbb{E}_\pi[Q]$.

$GI/GI/1$ analysis: "smoothed" quadratic test function. Now we consider upper bounding the mean queue length of the $GI/GI/1$ queue. As argued in Section 3.2, it suffices to focus on modified $GI/GI/1$ queue. In the modified $GI/GI/1$ queue, the stationarity of the test function Q^2 does not lead to a formula for $\mathbb{E}_\pi[Q]$, because Q no longer has a constant drift. Instead, the jumps of Q are state-dependent, modulated by the residual arrival time $R_a(t)$ and the residual completion time $R_s(t)$.

To reduce the problem to the familiar case of constant drift, one can consider the test function

$$f(R_s, R_a, Q) = (Q + \mu R_s - \lambda R_a)^2.$$

Intuitively, the additional terms $\mu R_s - \lambda R_a$ *smooths out* the dynamics of Q, so that $Q + \mu R_s - \lambda R_a$ has the constant drift $\lambda - \mu$ except when $Q = 0$. In particular, when a job arrives, $Q + \mu R_s - \lambda R_a$ increases by $1 - \lambda A$, and the amount of increase has a zero expectation; when a job completes and $Q > 0$, the change in $Q + \mu R_s - \lambda R_a$ again has a zero expectation.

The stationarity of this "smoothed" quadratic test function $(Q + \mu R_s - \lambda R_a)^2$ implies the following:

$$\begin{aligned}(1-\rho)\mathbb{E}_\pi[Q] = {} & \frac{\rho \mathrm{Var}[\lambda A] + \mathrm{Var}[\mu S] + 1 - \rho}{2} \\ & + \mathbb{E}_s[\mathbb{1}\{Q=0\}(\mu R_s - \lambda R_a)] - (1-\rho)\mathbb{E}_\pi[\mu R_s - \lambda R_a],\end{aligned} \tag{11}$$

where $\mathbb{E}_s[\cdot]$ denotes the expected value observed at a random completion event, to be formalized in Section 6.2. Here, the cross term $\mathbb{E}_s[\mathbb{1}\{Q=0\}(\mu R_s - \lambda R_a)]$ again comes from the fact that the drift of $Q + \mu R_s - \lambda R_a$ depends on whether $Q = 0$. Fortunately, this boundary effect is mild; in fact, we can upper bound this cross term by zero because $R_s = 0$ at any completion event. Thus, we obtain the upper bound $\mathbb{E}_\pi[Q] \le \big(\rho \mathrm{Var}[\lambda A] + \mathrm{Var}[\mu S] + 1-\rho\big)/\big(2(1-\rho)\big)$.

The above argument shall be formalized by the BAR framework, which we set up in Section 6.2. The technique of using a "smoothed" test function in BAR is invented by Miyazawa [62], and has been applied to various queueing systems with general interarrival-time and service-time distributions [see, e.g., 11, 14, 15, 19, 20, 22, 32, 33, 37, 62].

$GI/GI/n$ analysis: asymptotic independence. To upper bound the mean queue length of the $GI/GI/n$ queue, we consider the modified $GI/GI/n$ queue, and use the test function that straightforwardly generalizes the smoothed quadratic test function for modified $GI/GI/1$:

$$f(R_{s,1},\dots,R_{s,n},R_a,Q) = \Big(Q + \mu\sum_{j=1}^n R_{s,j} - \Lambda R_a\Big)^2.$$

As we show in Lemma 7, applying BAR to this test function yields

$$(1-\rho)\mathbb{E}_\pi[Q] = \frac{\rho\mathrm{Var}[\Lambda A] + \mathrm{Var}[\mu S] + 1-\rho}{2} + \sum_{j=1}^n \Gamma_{s,j} - \Gamma_a\,, \tag{12}$$

where $\Gamma_{s,j}$ and Γ_a are given by

$$\Gamma_{s,j} = \mathbb{E}_s\big[\mathbb{1}\{Q=0\}\mu R_{s,j}\big] - (1-\rho)\mathbb{E}_\pi\big[\mu R_{s,j}\big] \quad \forall j\in[n] \tag{13}$$

$$\Gamma_a = \mathbb{E}_s[\mathbb{1}\{Q=0\}\Lambda R_a] - (1-\rho)\mathbb{E}_\pi[\Lambda R_a]. \tag{14}$$

Because $\mathbb{E}_s[\mathbb{1}\{Q=0\}] = 1-\rho$, one can roughly view $\Gamma_{s,j}$ (or Γ_a) as the *covariance* between $\mathbb{1}\{Q=0\}$ and the residual service time $R_{s,j}$ (or the residual arrival time R_a). As in the case of $GI/GI/1$, these covariances arise from the fact that drift of the $Q + \mu\sum_{j=1}^n R_{s,j} - \Lambda R_a$ is not constant and depends on whether $Q = 0$; but unlike $GI/GI/1$, now there are $(n+1)$ covariance terms rather than 2. This manifests the complexity caused by multiple servers. Equations similar to (12) have been used explicitly or implicitly in many prior studies of the $GI/GI/n$ queue and related models, where the steady-state mean queue length or workload is expressed in terms of $\mathrm{Var}[\Lambda A]$, $\mathrm{Var}[\mu S]$, and similar covariance terms [31, 36, 38, 53, 72]. However, none of these papers can derive an $O\big(1/(1-\rho)\big)$ bound valid for all $\rho < 1$ and n through such equations, despite the conjecture to that effect [53, Conjecture 1].

The core technical challenge of this paper is to show that the sum of the covariances

$$\sum_{j=1}^n \Gamma_{s,j} = \sum_{j=1}^n \Big(\mathbb{E}_s\big[\mathbb{1}\{Q=0\}\mu R_{s,j}\big] - (1-\rho)\mathbb{E}_\pi\big[\mu R_{s,j}\big]\Big)$$

remains bounded as $n\to\infty$, or equivalently, $\Gamma_{s,j} = O(1/n)$ for each $j\in[n]$. This will imply the desired $O\big(1/(1-\rho)\big)$ upper bound once combined with (12). Unlike the case for $GI/GI/1$, we can no longer argue that $R_{s,j} = 0$ under the expectation $\mathbb{E}_s[\cdot]$, given that there are multiple servers.

Our idea is to show that $\mathbb{1}\{Q=0\}$ and the residual service time of the j-th server $R_{s,j}$ are *asymptotically independent*, so that their covariance $\Gamma_{s,j}$ diminishes at the rate $O(1/n)$ as $n\to\infty$. Why should $\mathbb{1}\{Q=0\}$ and $R_{s,j}$ be nearly independent, given that the queue length is constantly interacting with the j-th server's completion process? Intuitively, the queue length Q in a modified $GI/GI/n$ queue is the joint effect of n independent service-completion processes, so each process should only have $1/n$ fraction of "responsibility" for the realization of the random variable $\mathbb{1}\{Q=0\}$. To capture the effect of the j-th server's completion process on the queue length, we ask the following motivating question:

How would the queue length change if the j-th server had not existed?

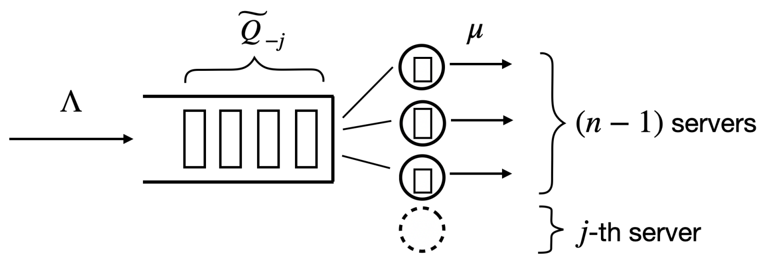

Fig 2: An illustration of the j-th leave-one-out system, where the j-th server does not take jobs from the queue.

An attempt to answer this question inspires a simple proof of the asymptotic independence. We consider the counterfactual queue length $\widetilde{Q}_{-j}$ if j-th server had not existed in the modified $GI/GI/n$ queue (see Section 6.1 for the precise definition). Crucially, the indicator $\mathbb{1}\{\widetilde{Q}_{-j}=0\}$ serves as a "bridge" between $\mathbb{1}\{Q=0\}$ and $R_{s,j}$: it is independent of $R_{s,j}$ yet closely approximates $\mathbb{1}\{Q=0\}$. This allows us to show that $\mathbb{1}\{Q=0\}$ and $R_{s,j}$ are asymptotically independent. The detailed argument can be found in the proof of Lemma 8 in Section 7.2.

6. Technical framework and supporting lemmas. In this section, we formally set up the technical framework and state the supporting lemmas for Theorem 1. We define the leave-one-out systems in Section 6.1, set up the Basic Adjoint Relationship in Section 6.2, prove some additional supporting lemmas in Section 6.3.

To simplify the presentation, we adopt the following assumption. Our results hold without this assumption, and we will outline the necessary modifications for the general case in Remark 2.

ASSUMPTION 4. In the modified $GI/GI/n$ queue, no two events (i.e., an arrival and a completion, two arrivals, or two completions) happen at the same time.

6.1. *Leave-one-out systems.* In this section, we construct a set of auxiliary systems, referred to as *leave-one-out systems*. Each leave-one-out system captures the counterfactual queue length if a certain server were removed from the main n-server modified $GI/GI/n$ system under study (referred to as the "main system" or simply "the modified $GI/GI/n$ queue" from now on). Leave-one-out systems play a crucial role in our proof of Theorem 1 in Section 7, representing a key innovation of this paper.

For each $j\in[n]$, the j-th leave-one-out system is a modified $GI/GI/(n-1)$ system coupled with the main system; it shares the arrival events and the completion events of all but the j-th server with the main system, as illustrated in Figure 2. In other words, each job arrival in the main system also increases $\widetilde{Q}_{-j}(t)$ by 1 for all $j\in[n]$, and each completion of the i-th server reduces $\widetilde{Q}_{-j}(t)$ by $\mathbb{1}\{\widetilde{Q}_{-j}(t-)>0\}$ for $j\in[n]\backslash i$. We initialize $\widetilde{Q}_{-j}(0)=Q(0)$ all $j\in[n]$, and denote the joint queue lengths of all leave-one-out systems as $\widetilde{Q}(t)\triangleq(\widetilde{Q}_{-j}(t))_{j\in[n]}$.

As one would expect, the queue length of each leave-one-out system pathwise dominates the queue length of the main system:

LEMMA 1. *For each $j=1,2,\dots n$, the queue length of the j-th leave-one-out system pathwise upper bounds the queue length of the main system, i.e.,*

$$\widetilde{Q}_{-j}(t)\geq Q(t)\quad\forall t\geq 0.$$

PROOF. We define the stopping time $\tau = \inf\{t \geq 0 \colon \widetilde{Q}_{-j}(t) \leq Q(t) - 1\}$. It suffices to show that $\tau = \infty$ for all sample paths. First, because $\widetilde{Q}_{-j}(0) = Q(0)$, by the right-continuity of $\widetilde{Q}_{-j}(t)$ and $Q(t)$, we must have $\tau > 0$. Suppose that there exists a sample path such that $\tau < \infty$, then we have $\widetilde{Q}_{-j}(\tau-) = Q(\tau-)$ and $\widetilde{Q}_{-j}(\tau) \leq Q(\tau) - 1$. This implies that at time τ, there must be at least a completion event that affects the j-th leave-one-out system but not the main system, or an arrival event that affects the main system but not the j-th leave-one-out system. However, both cases are impossible, which is a contradiction. Consequently, for all sample paths, $\tau = \infty$, so $\widetilde{Q}_{-j}(t) \geq Q(t)$ for all $t \geq 0$. □

Finally, note that by Assumption 3, each j-th leave-one-out system is positive Harris recurrent when $\rho < 1 - 1/n$.

6.2. *Basic Adjoint Relationship.* As outlined in Section 5, our analysis follows the framework of Basic Adjoint Relationship (BAR) to perform the steady-state analysis. In this section, we formally set up the preliminaries of BAR, following Section 6 of [14].

We first define some basic notation. We focus on analyzing the Markov process $\overline{X}(t)$, which is either $X(t)$ if $1 - 1/n \leq \rho < 1$ or $(X(t), \widetilde{Q}(t))$ if $\rho < 1 - 1/n$. Let the state space of $\overline{X}(t)$ be $\overline{\mathcal{X}}$, which is $\mathbb{R}_{\geq 0}^{n+1} \times \mathbb{Z}_{\geq 0}$ or $\mathbb{R}_{\geq 0}^{n+1} \times \mathbb{Z}_{\geq 0}^{n+1}$, depending on ρ. By Assumption 3, in both cases, $\overline{X}(t)$ has a unique stationary distribution π, capturing its long-run time-average behavior. From now on until the end of Section 7, we slightly abuse the notation to let $\overline{X}(0) \sim \pi$, which implies $\overline{X}(t) \sim \pi$ for all $t \geq 0$.

On a high level, BAR is based on the fact that any function of the state has zero expected changes in the steady state. In our case, this implies that the joint effects of the arrivals, completions, and continuous decrease of the residual times on the function's value should be zero in the steady state. In the following, we define the concepts that help quantify these three types of changes in function value.

We define the differential operator $\mathcal{A}$ as

$$\mathcal{A} f(\bar{x}) \triangleq - \sum_{\ell=1}^{n+1} \partial_\ell^+ f(\bar{x}) \quad \forall \bar{x} \in \overline{\mathcal{X}}, \tag{15}$$

where ∂_ℓ^+ denotes the right partial derivative with respect to the ℓ-th coordinate, and $f \colon \overline{\mathcal{X}} \to \mathbb{R}$ is any function that has right partial derivatives with respect to the first $(n+1)$ coordinates (the coordinates corresponding to $R_{s,i}$'s and R_a). This operator captures the instantaneous changing rate of f in the state $\bar{x}$ caused by the decrease of the residual times; the corresponding steady-state changing rate is thus given by $\mathbb{E}_\pi\left[\mathcal{A} f(\overline{X})\right]$.

Next, we consider the changes in $f(X(t))$ caused by the arrivals. Since arrivals occur at discrete points, we cannot define their rates. Instead, we use the so-called *Palm distributions*, which are probability measures over $\overline{\mathcal{X}}^2$ and captures the probability of the pair of states before and after a random jump in the steady state. Formally, the *Palm distribution at arrivals* $\mathbb{P}_a$ is defined as

$$\mathbb{P}_a[B] = \frac{1}{\Lambda h} \mathbb{E}\left[\sum_{k=1}^{\infty} \mathbb{1}\left\{\left(\overline{X}(\tau_a^k), \overline{X}(\tau_a^k -)\right) \in B\right\} \mathbb{1}\left\{\tau_a^k \leq h\right\}\right], \tag{16}$$

where B is any measurable subset of $\overline{\mathcal{X}}^2$ and τ_a^k is the time of the k-th arrival event; h is any positive number, and $\mathbb{P}_a$ does not depend on the choice of h due to the stationarity of $\overline{X}(t)$ [2, Theorem VII.6.3]. Roughly speaking, $\mathbb{P}_a$ represents the long-run average distribution evaluated at arrival events; this can be seen by taking a large h, and note that Λ is the long-term average frequency of arrival events. We refer the readers to [2, Section VII.6] or [14, Section 6] for a detailed background of Palm distributions induced by point processes.

Given the Palm distribution $\mathbb{P}_a$, we can define the corresponding Palm expectation $\mathbb{E}_a[\cdot]$: For any measurable function $g : \overline{\mathcal{X}}^2 \to \mathbb{R}$, we have

$$\mathbb{E}_a\left[g\left(\overline{X}^+, \overline{X}^-\right)\right] \triangleq \frac{1}{\Lambda h}\mathbb{E}\left[\sum_{k=1}^{\infty} g(\overline{X}(\tau_a^k), \overline{X}(\tau_a^k-))\mathbb{1}\{\tau_a^k \le h\}\right], \tag{17}$$

where $(\overline{X}^+, \overline{X}^-) \in \overline{\mathcal{X}}^2$ is a pair of dummy variables following the distribution $\mathbb{P}_a$, and h is any positive number. Then for any measurable function $f : \overline{\mathcal{X}} \to \mathbb{R}$, its long-run average rate of changes caused by arrivals can be represented by $\Lambda\mathbb{E}_a\left[f(\overline{X}^+) - f(\overline{X}^-)\right]$.

We similarly define $\mathbb{P}_{s,i}$ and $\mathbb{E}_{s,i}[\cdot]$ for the completion events of the i-th server, for $i \in [n]$.

Now we are ready to state the main lemma for BAR, which connects the stationary distribution π with the set of Palm distributions $\mathbb{P}_a$ and and $\mathbb{P}_{s,i}$ defined above. The proof of Lemma 2 is provided in Section D.2, which is adapted from [14, Section 6.1].

LEMMA 2 (Basic Adjoint Relationship for $GI/GI/n$). *Consider the modified $GI/GI/n$ system satisfying Assumptions 1, 2 and 3, with load $\rho < 1$. For any function $f : \overline{\mathcal{X}} \to \mathbb{R}$ that is bounded, locally Lipschitz continuous, and partially differentiable from the right with respect to its first $(n+1)$ coordinates, we have*

$$\mathbb{E}_\pi\left[\mathcal{A}f(\overline{X})\right] + \Lambda\mathbb{E}_a\left[f(\overline{X}^+) - f(\overline{X}^-)\right] + \mu\sum_{i=1}^{n}\mathbb{E}_{s,i}\left[f(\overline{X}^+) - f(\overline{X}^-)\right] = 0, \tag{18}$$

where $\overline{X}, \overline{X}^+, \overline{X}^- \in \overline{\mathcal{X}}$ are random elements following the distributions specified in the expectation subscripts. Moreover, the Palm distributions $\mathbb{P}_a$ and $\mathbb{P}_{s,i}$ satisfy the following with probability 1:

(i) *If $(\overline{X}^+, \overline{X}^-) \sim \mathbb{P}_a$, then $R_a^- = 0$ and the coordinates of $\overline{X}^+$ and $\overline{X}^-$ satisfy the relations:*

$$R_a^+ = R_a^- + A \tag{19}$$

$$Q^+ = Q^- + 1 \tag{20}$$

$$\widetilde{Q}_{-j}^+ = \widetilde{Q}_{-j}^- + 1 \quad \forall j \in [n] \quad \textit{if } \rho < 1 - 1/n, \tag{21}$$

where A represents a generic interarrival time and is independent of $\overline{X}^-$; other coordinates of $\overline{X}^+$ are identical to $\overline{X}^-$. [1]

(ii) *For each $i \in [n]$, if $(\overline{X}^+, \overline{X}^-) \sim \mathbb{P}_{s,i}$, then $R_{s,i}^- = 0$ and the coordinates of $\overline{X}^+$ and $\overline{X}^-$ satisfy the relations:*

$$R_{s,i}^+ = R_{s,i}^- + S \tag{22}$$

$$Q^+ = Q^- - \mathbb{1}\{Q^- > 0\} \tag{23}$$

$$\widetilde{Q}_{-j}^+ = \widetilde{Q}_{-j}^- - \mathbb{1}\left\{\widetilde{Q}_{-j}^- > 0\right\} \quad \forall j \in [n]\backslash\{i\} \quad \textit{if } \rho < 1 - 1/n, \tag{24}$$

where S represents a generic service time and is independent of $\overline{X}^-$; other coordinates of $\overline{X}^+$ are identical to $\overline{X}^-$.

Next, we define some convenient notational shorthand. We define the Palm distribution $\mathbb{P}_s[\cdot] \triangleq \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}_{s,i}[\cdot]$ and the corresponding expectation $\mathbb{E}_s\left[g(\overline{X}^+, \overline{X}^-)\right] \triangleq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{s,i}\left[g(\overline{X}^+, \overline{X}^-)\right]$. Under this notation, (18) can be compactly rewritten as

$$\mathbb{E}_\pi\left[\mathcal{A}f(\overline{X})\right] + \Lambda\mathbb{E}_a\left[f(\overline{X}^+) - f(\overline{X}^-)\right] + n\mu\mathbb{E}_s\left[f(\overline{X}^+) - f(\overline{X}^-)\right] = 0. \tag{25}$$

[1] Any superscript applied to an element of $\overline{\mathcal{X}}$ (e.g., $\overline{X}^-$ or $\overline{X}^+$) propagates to all of its coordinates.

We also adopt the simplification that, when a Palm expectation only involves the pre-event state, $\overline{X}^-$, we omit its superscript and simply write $\overline{X}$.

REMARK 2. We have adopted the assumption that no events occur simultaneously (Assumption 4) to simplify the presentation. However, our results still hold without this assumption, and the necessary changes are outlined as follows.

The main changes are in the definitions of the Palm distributions. When m events happen at time t, we use $\overline{X}(t,0), \overline{X}(t,1), \ldots \overline{X}(t,m)$ to denote the intermediate states after sequentially applying the state-updating subroutines for each type of event involved (see Section 3.2); in particular, $\overline{X}(t,0) \triangleq \overline{X}(t-)$ and $\overline{X}(t,m) \triangleq \overline{X}(t)$. Let (τ_a^k, m_a^k) be the k-th (t,m)-pair such that $\overline{X}(t,m)$ is an intermediate state before an arrival update. Then the definition of $\mathbb{P}_a$ should be generalized as

$$\mathbb{P}_a[B] = \frac{1}{\Lambda h}\mathbb{E}\left[\sum_{k=1}^{\infty} \mathbb{1}\left\{\left(\overline{X}(\tau_a^k, m_a^k+1), \overline{X}(\tau_a^k, m_a^k)\right) \in B\right\}\mathbb{1}\left\{\tau_a^k \le h\right\}\right]. \tag{26}$$

The Palm distributions $\mathbb{P}_{s,i}$ for $i \in [n]$ should be generalized analogously. Intuitively, $\mathbb{P}_a$ and $\mathbb{P}_{s,i}$ still represent the average state distributions observed by arrival events or completion events, although the states at some time points could be counted multiple times or involved in multiple Palm distributions. The same definition has also been used in [14] to deal with simultaneous events.

With the new definitions of $\mathbb{P}_a$ and $\mathbb{P}_{s,i}$, the statement of Lemma 2 is valid without requiring any other changes. The other lemmas, including those for heterogenous $GI/GI/n$ in Section 8, also hold verbatim. We leave the detailed verifications to readers.

Finally, we mention an interesting detail. As we can see in Section 3.2 when constructing the modified $GI/GI/n$, the order of updating the state in the presence of simultaneous events could affect the system dynamics. However, our generalized proof should work for any order of updates. Intuitively, whether we apply the arrival subroutine before or after a completion subroutine would affect the distribution of the distribution of R_a^- under $\mathbb{P}_{s,i}$, but not those facts stated in Lemma 2.

6.3. *Other supporting lemmas.* In this section, we state some supporting lemmas for the proof of Theorem 1. These lemmas are about some intuitive facts, so we briefly comment on their intuitions and leave their proofs to Section D.

Recall that by our previous assumptions, conditioned on any age, the mean residual service time is upper-bounded by $R_s^{\max}$ (Assumption 1), and the mean residual arrival time is lower-bounded by $R_a^{\min}$. Lemma 3 below further shows these bounds on the conditional expectations remain true after conditioning on the queue lengths Q and $\widetilde{Q}$. This is because the age of the service time or interarrival time is effectively a sufficient statistic of the history of arrival and completion events for inferring the residual times. Since the queue lengths are also determined by this history, they provide no additional information relevant to the residual times.

LEMMA 3 (Bounded mean residual times). *Consider the modified $GI/GI/n$ queue satisfying Assumptions 1 and 3, with the load $\rho < 1$. If $\rho < 1 - 1/n$, for any $j \in [n]$,*

$$\mu\mathbb{E}_\nu\left[R_{s,j} \,\middle|\, Q, \widetilde{Q}\right] \le R_s^{\max} \tag{27}$$

$$\Lambda\mathbb{E}_\nu\left[R_a \,\middle|\, Q, \widetilde{Q}\right] \ge R_a^{\min} \tag{28}$$

where $\mathbb{E}_\nu[\cdot]$ can denote either $\mathbb{E}_\pi[\cdot]$ or $\mathbb{E}_{s,i}[\cdot]$ for $i \ne j$; if $1 - 1/n \le \rho < 1$, we have analogs of (27) and (28) without conditioned on $\widetilde{Q}$.

The following lemma establishes that $R_{s,j}$ and $\widetilde{Q}_{-j}$ are uncorrelated under the Palm distribution $\mathbb{P}_{s,i}$ for $i \neq j$. As the system under study is driven by mutually independent renewal processes, the intuition for this lemma is straightforward: $R_{s,j}$ depends only on the server j's completion process, whereas $\widetilde{Q}_{-j}$ and $\mathbb{P}_{s,i}$ are determined by the set of renewal processes that exclude that of server j's. The formal proof also involves a mixing argument of the completion processes, which relies on S being non-lattice (Assumption 2). Notably, this lemma is the only part of our analysis that requires the non-lattice assumption.

LEMMA 4 (Uncorrelation lemma). *Consider the modified $GI/GI/n$ queue satisfying Assumptions 1, 2, and 3, with the load $\rho < 1 - 1/n$. For any $i, j \in [n]$ such that $i \neq j$, we have*

$$\mathbb{E}_{s,i}\left[\mathbb{1}\left\{\widetilde{Q}_{-j} = 0\right\} R_{s,j}\right] = \mathbb{P}_{s,i}\left[\widetilde{Q}_{-j} = 0\right] \mathbb{E}_{\pi}\left[R_{s,j}\right]. \tag{29}$$

The next lemma shows several basic equalities about the residual times of the renewal processes and the idle probability of the queue. These equalities directly follow from BAR and are proved in Section D.5.

LEMMA 5. *Consider the modified $GI/GI/n$ queue satisfying Assumptions 1 and 3, with the load $\rho < 1$. For $i, j \in [n]$ such that $i \neq j$, we have*

$$\mathbb{E}_{\pi}\left[R_{s,i}\right] = \frac{1}{2}\mu\mathbb{E}\left[S^2\right] \tag{30}$$

$$\mathbb{E}_{\pi}\left[R_a\right] = \frac{1}{2}\Lambda\mathbb{E}\left[A^2\right] \tag{31}$$

$$\mathbb{E}_{s,i}\left[R_{s,j}\right] + \mathbb{E}_{s,j}\left[R_{s,i}\right] = \mu\mathbb{E}\left[S^2\right] \tag{32}$$

$$\mathbb{E}_{s,i}\left[R_a\right] + \mathbb{E}_a\left[R_{s,i}\right] = \frac{1}{2}\mu\mathbb{E}\left[S^2\right] + \frac{1}{2}\Lambda\mathbb{E}\left[A^2\right] \tag{33}$$

$$\mathbb{P}_s\left[Q = 0\right] = 1 - \rho \tag{34}$$

$$\frac{1}{n}\sum_{i:\, i \neq j} \mathbb{P}_{s,i}\left[\widetilde{Q}_{-j} = 0\right] = 1 - \rho - \frac{1}{n} \quad \text{if } \rho < 1 - 1/n. \tag{35}$$

Next, we show that the queue length has finite means under the Palm distributions $\mathbb{P}_a$ and $\mathbb{P}_{s,i}$, utilizing the assumption that it has a finite mean under the stationary distribution π (Assumption 3).

LEMMA 6. *Consider the modified $GI/GI/n$ system satisfying Assumptions 1 and 3, with the load $\rho < 1$. Then we have $\mathbb{E}_a[Q] < \infty$ and $\mathbb{E}_{s,i}[Q] < \infty$ for all $i \in [n]$.*

7. Proof of main theorem. This section is dedicated to proving Theorem 1, following the two-step approach outlined in Section 5:

Step 1 (Section 7.1): We derive an exact expression for the mean queue length, $\mathbb{E}_{\pi}[Q]$. This expression connects $\mathbb{E}_{\pi}[Q]$ to the load ρ, the variances $\mathrm{Var}[\mu S]$ and $\mathrm{Var}[\Lambda A]$, and some "covariance" terms.

Step 2 (Section 7.2): We bound these "covariance" terms.

The first step follows from the established framework of BAR, while the second step contains the novel argument based on our leave-one-out technique. Substituting the bound in Step 2 into the expression in Step 1 proves Theorem 1.

7.1. *Step 1: exact expression of* $\mathbb{E}[Q]$.

LEMMA 7. *Consider the modified $GI/GI/n$ system satisfying Assumptions 1 and 3, with the load $\rho < 1$. We have*

$$(1-\rho)\mathbb{E}_\pi[Q] = \frac{\rho \mathrm{Var}[\Lambda A] + \mathrm{Var}[\mu S] + 1 - \rho}{2} + \sum_{j=1}^{n} \Gamma_{s,j} - \Gamma_a \,, \tag{36}$$

where $\Gamma_{s,j}$ and Γ_a are defined as:

$$\Gamma_{s,j} = \mathbb{E}_s\left[\mathbb{1}\{Q=0\}\mu R_{s,j}\right] - (1-\rho)\mathbb{E}_\pi\left[\mu R_{s,j}\right] \quad \forall j \in [n] \tag{37}$$

$$\Gamma_a = \mathbb{E}_s\left[\mathbb{1}\{Q=0\}\Lambda R_a\right] - (1-\rho)\mathbb{E}_\pi\left[\Lambda R_a\right], \tag{38}$$

and recall that $\mathbb{E}_s[\cdot] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{s,i}[\cdot]$ is the Palm expectation with respect to all completion events.

REMARK 3. When n is large, the term $\Gamma_{s,j}$ can be roughly viewed as the covariance between $\mathbb{1}\{Q=0\}$ and $\mu R_{s,j}$ under the measure $\mathbb{P}_s$, because $\mathbb{P}_s[Q=0] = 1-\rho$ and $\mathbb{E}_s\left[R_{s,j}\right] = \frac{1}{n}\sum_{i:\, i\neq j}\mathbb{E}_{s,i}\left[R_{s,j}\right] = \frac{n-1}{n}\mathbb{E}_\pi\left[R_{s,j}\right]$. Here we use the fact that $\mathbb{E}_{s,i}\left[R_{s,j}\right] = \mathbb{E}_\pi\left[R_{s,j}\right]$, which is slightly stronger than (32) in Lemma 5, but can be shown by replacing $\mathbb{1}\left\{\widetilde{Q}_{-j} = 0\right\}$ with 1 in the proof of Lemma 4. We skip the detailed argument because it does not affect our results.

The term Γ_a can be interpreted similarly if the interarrival-time distribution is non-lattice.

PROOF OF LEMMA 7. For any integer $\kappa > 0$, we consider the test function

$$f(\overline{X}) = \left(Q \wedge \kappa + \mu \sum_{j=1}^{n} R_{s,j} \wedge \kappa - \Lambda R_a \wedge \kappa\right)^2,$$

which is a truncated version of the quadratic function $\left(Q + \mu\sum_{j=1}^{n} R_{s,j} - \Lambda R_a\right)^2$. Since f is bounded, locally Lipschitz continuous, and has right partial derivatives with respect to R_a and $R_{s,i}$'s, Basic Adjoint Relationship (2) implies that

$$\begin{aligned}
0 = {} & 2\mathbb{E}_\pi\left[\left(Q \wedge \kappa + \mu\sum_{j=1}^{n} R_{s,j} \wedge \kappa - \Lambda R_a \wedge \kappa\right)\left(-\mu\sum_{j=1}^{n}\mathbb{1}\left\{R_{s,j} \le \kappa\right\} + \Lambda \mathbb{1}\{R_a \le \kappa\}\right)\right] \\
& + \Lambda \mathbb{E}_a\left[\left(Q \wedge \kappa + \mu\sum_{j=1}^{n} R_{s,j} \wedge \kappa + \mathbb{1}\{Q \le \kappa - 1\} - \Lambda A \wedge \kappa\right)^2 - \left(Q \wedge \kappa + \mu\sum_{j=1}^{n} R_{s,j} \wedge \kappa\right)^2\right] \\
& + n\mu\mathbb{E}_s\left[\left(Q \wedge \kappa + \mu\sum_{j=1}^{n} R_{s,j} \wedge \kappa - \Lambda R_a \wedge \kappa - \mathbb{1}\{0 < Q \le \kappa\} + \mu S \wedge \kappa\right)^2\right] \\
& - n\mu\mathbb{E}_s\left[\left(Q \wedge \kappa + \mu\sum_{j=1}^{n} R_{s,j} \wedge \kappa - \Lambda R_a \wedge \kappa\right)^2\right].
\end{aligned}$$

Note that to get the $\mathbb{E}_a[\cdot]$ term above, we have used the fact that $R_a = 0$ with probability 1 under $\mathbb{P}_a$. Expanding the squares and rearranging the terms, the above equation simplifies into

$$2\mathbb{E}_\pi\left[\left(Q \wedge \kappa + \mu\sum_{j=1}^{n} R_{s,j} \wedge \kappa - \Lambda R_a \wedge \kappa\right)\left(\mu\sum_{j=1}^{n}\mathbb{1}\left\{R_{s,j} \le \kappa\right\} - \Lambda \mathbb{1}\{R_a \le \kappa\}\right)\right]$$

$$= \Lambda \mathbb{E}_a\left[2\Big(Q \wedge \kappa + \mu \sum_{j=1}^{n} R_{s,j} \wedge \kappa\Big)\Big(\mathbb{1}\{Q \le \kappa - 1\} - \Lambda A \wedge \kappa\Big) + \Big(\mathbb{1}\{Q \le \kappa - 1\} - \Lambda A \wedge \kappa\Big)^2\right]$$

$$+ n\mu \mathbb{E}_s\left[2\Big(Q \wedge \kappa + \mu \sum_{j=1}^{n} R_{s,j} \wedge \kappa - \Lambda R_a \wedge \kappa\Big)\Big(-\mathbb{1}\{0 < Q \le \kappa\} + \mu S \wedge \kappa\Big)\right]$$

$$+ n\mu \mathbb{E}_s\left[\Big(-\mathbb{1}\{0 < Q \le \kappa\} + \mu S \wedge \kappa\Big)^2\right].$$

Now we let $\kappa \to \infty$ in the equality and attempt to apply the dominated convergence theorem. To do this, recall the following finiteness results: $\mathbb{E}_\pi[Q] < \infty$ (Assumption 3), $\mathbb{E}_a[Q], \mathbb{E}_s[Q] < \infty$ (Lemma 6), $\mathbb{E}_\pi[R_a], \mathbb{E}_{s,i}[R_a], \mathbb{E}_\pi\left[R_{s,i}\right], \mathbb{E}_a\left[R_{s,i}\right], \mathbb{E}_{s,i}\left[R_{s,j}\right] < \infty$ (Lemma 5), $\mathbb{E}_a[R_a] = \mathbb{E}_{s,i}\left[R_{s,i}\right] = 0$, and $\mathbb{E}\left[A^2\right], \mathbb{E}\left[S^2\right] < \infty$ (Assumption 1). Then in the last displayed equality, it is not hard to see that the expression inside $\mathbb{E}_\pi[\cdot]$ is dominated by $(Q + \sum_{j=1}^{n} \mu R_{s,j} + \Lambda R_a)(n\mu + \Lambda)$, which is independent of κ and has a finite mean under the measure π. Similarly, one can see that the expressions inside $\mathbb{E}_a[\cdot]$ and $\mathbb{E}_s[\cdot]$ are dominated by expressions that have finite means under $\mathbb{P}_a$ and $\mathbb{P}_s$, respectively. Therefore, the dominated convergence theorem implies that

(39)
$$2(n\mu - \Lambda)\mathbb{E}_\pi\left[Q + \mu \sum_{j=1}^{n} R_{s,j} - \Lambda R_a\right]$$

(40)
$$= \Lambda \mathbb{E}_a\left[2\Big(Q + \mu \sum_{j=1}^{n} R_{s,j}\Big)\Big(1 - \Lambda A\Big) + \Big(1 - \Lambda A\Big)^2\right]$$

(41)
$$+ n\mu \mathbb{E}_s\left[2\Big(Q + \mu \sum_{j=1}^{n} R_{s,j} - \Lambda R_a\Big)\Big(-\mathbb{1}\{Q > 0\} + \mu S\Big) + \Big(-\mathbb{1}\{Q > 0\} + \mu S\Big)^2\right].$$

Next, we simplify the terms in (40), which correspond to the arrival events:

$$\Lambda \mathbb{E}_a\left[2\Big(Q + \mu \sum_{j=1}^{n} R_{s,j}\Big)\Big(1 - \Lambda A\Big) + \Big(1 - \Lambda A\Big)^2\right]$$

$$= 2\Lambda \mathbb{E}_a\left[Q + \mu \sum_{j=1}^{n} R_{s,j}\right]\mathbb{E}\Big[1 - \Lambda A\Big] + \Lambda \mathbb{E}\Big[(1 - \Lambda A)^2\Big]$$

$$= \Lambda \mathrm{Var}\left[\Lambda A\right], \tag{42}$$

where the first equality is due to the fact that A is independent of Q and $R_{s,j}$; the second equality is due to $\Lambda \mathbb{E}[A] = 1$.

We then simplify the terms in (41), which correspond to the completion events:

$$n\mu \mathbb{E}_s\left[2\Big(Q + \mu \sum_{j=1}^{n} R_{s,j} - \Lambda R_a\Big)\Big(-\mathbb{1}\{Q > 0\} + \mu S\Big) + \Big(-\mathbb{1}\{Q > 0\} + \mu S\Big)^2\right]$$

$$
\begin{aligned}
&= 2n\mu\mathbb{E}_s\left[Q + \mu\sum_{j=1}^n R_{s,j} - \Lambda R_a\right]\mathbb{E}\Big[-1+\mu S\Big] \\
&\quad + 2n\mu\mathbb{E}_s\left[\Big(Q + \mu\sum_{j=1}^n R_{s,j} - \Lambda R_a\Big)\mathbb{1}\{Q=0\}\right] \\
&\quad + n\mu\mathbb{E}\left[(-1+\mu S)^2\right] + 2n\mu\mathbb{P}_s\left[Q=0\right]\mathbb{E}\left[-1+\mu S\right] + n\mu\mathbb{P}_s\left[Q=0\right] \\
&= 2n\mu\mathbb{E}_s\left[\Big(\mu\sum_{j=1}^n R_{s,j} - \Lambda R_a\Big)\mathbb{1}\{Q=0\}\right] + n\mu\mathrm{Var}\left[\mu S\right] + n\mu\mathbb{P}_s\left[Q=0\right] \\
&= 2n\mu\mathbb{E}_s\left[\Big(\mu\sum_{j=1}^n R_{s,j} - \Lambda R_a\Big)\mathbb{1}\{Q=0\}\right] + n\mu\mathrm{Var}\left[\mu S\right] + n\mu(1-\rho), \qquad (43)
\end{aligned}
$$

where the first equality uses $\mathbb{1}\{Q>0\} = 1 - \mathbb{1}\{Q=0\}$ and the fact that S is independence of Q, $R_{s,j}$, and R_a; the second equality is due to $\mathbb{E}\left[\mu S\right] = 1$; the third equality is due to $\mathbb{P}_s\left[Q=0\right] = 1-\rho$ (Lemma 5).

Substituting (42) and (43) back to (40) and (41), we get

$$
\begin{aligned}
&2(n\mu-\Lambda)\mathbb{E}_\pi\left[Q + \mu\sum_{j=1}^n R_{s,j} - \Lambda R_a\right] \\
&\quad = 2n\mu\mathbb{E}_s\left[\Big(\mu\sum_{j=1}^n R_{s,j} - \Lambda R_a\Big)\mathbb{1}\{Q=0\}\right] + \Lambda\mathrm{Var}\left[\Lambda A\right] + n\mu\mathrm{Var}\left[\mu S\right] + n\mu(1-\rho).
\end{aligned}
$$

Dividing both sides by $2n\mu$ and using the fact that $\Lambda/(n\mu) = \rho$, the last equality becomes

$$
\begin{aligned}
&(1-\rho)\mathbb{E}_\pi\left[Q + \mu\sum_{j=1}^n R_{s,j} - \Lambda R_a\right] \\
&\quad = \mathbb{E}_s\left[\Big(\mu\sum_{j=1}^n R_{s,j} - \Lambda R_a\Big)\mathbb{1}\{Q=0\}\right] + \frac{\rho\mathrm{Var}\left[\Lambda A\right] + \mathrm{Var}\left[\mu S\right] + 1-\rho}{2}.
\end{aligned}
$$

Rearranging the terms, we get

$$
\begin{aligned}
(1-\rho)\mathbb{E}_\pi\left[Q\right] &= \frac{\rho\mathrm{Var}\left[\Lambda A\right] + \mathrm{Var}\left[\mu S\right] + 1-\rho}{2} \\
&\quad + \mathbb{E}_s\left[\Big(\mu\sum_{j=1}^n R_{s,j} - \Lambda R_a\Big)\mathbb{1}\{Q=0\}\right] - (1-\rho)\mathbb{E}_\pi\left[\mu\sum_{j=1}^n R_{s,j} - \Lambda R_a\right],
\end{aligned}
$$

which implies (36). □

7.2. *Step 2: analyzing the covariance terms.*

LEMMA 8. *Consider the modified $GI/GI/n$ system satisfying Assumptions 1 and 3, with the load $\rho<1$. Let $\Gamma_{s,j}$ and Γ_a be the terms defined in (37) and (38), respectively. Then*

$$\Gamma_{s,j} \le \min\Big(1-\rho, \frac{1}{n}\Big)\Big(R_s^{\max} - \frac{1}{2}\mathbb{E}\left[(\mu S)^2\right]\Big) \qquad (44)$$

$$-\Gamma_a \le (1-\rho)\Big(\frac{1}{2}\mathbb{E}\left[(\Lambda A)^2\right] - R_a^{\min}\Big) \qquad (45)$$

PROOF. Recall that

$$\Gamma_{s,j} = \mathbb{E}_s\left[\mathbb{1}\{Q=0\}\mu R_{s,j}\right] - (1-\rho)\mathbb{E}_\pi\left[\mu R_{s,j}\right] \quad \forall j \in [n] \tag{37}$$

$$\Gamma_a = \mathbb{E}_s\left[\mathbb{1}\{Q=0\}\Lambda R_a\right] - (1-\rho)\mathbb{E}_\pi\left[\Lambda R_a\right], \tag{38}$$

We first bound $\Gamma_{s,j}$ and then bound $-\Gamma_a$.

Bounding $\Gamma_{s,j}$. We consider two cases based on whether $\rho \geq 1 - 1/n$.

Case 1: $1 - 1/n \leq \rho < 1$**.**

$$\begin{aligned}
\Gamma_{s,j} &= \mathbb{E}_s\left[\mathbb{1}\{Q=0\}\mu R_{s,j}\right] - (1-\rho)\mathbb{E}_\pi\left[\mu R_{s,j}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{s,i}\left[\mathbb{1}\{Q=0\}\mu R_{s,j}\right] - (1-\rho)\mathbb{E}_\pi\left[\mu R_{s,j}\right] \\
&\leq \frac{1}{n}\sum_{i:\, i\neq j} \mathbb{P}_{s,i}\left[Q=0\right] R_s^{\max} - (1-\rho)\mathbb{E}_\pi\left[\mu R_{s,j}\right] \qquad (46)\\
&\leq \mathbb{P}_s\left[Q=0\right] R_s^{\max} - (1-\rho)\mathbb{E}_\pi\left[\mu R_{s,j}\right] \\
&= (1-\rho) R_s^{\max} - (1-\rho)\mathbb{E}_\pi\left[\mu R_{s,j}\right] \qquad (47)\\
&= (1-\rho)\left(R_s^{\max} - \frac{1}{2}\mathbb{E}\left[(\mu S)^2\right]\right), \qquad (48)
\end{aligned}$$

where the inequality (46) is due to $R_{s,i} = 0$ with probability 1 under $\mathbb{P}_{s,i}$ (Lemma 2), and $\mathbb{E}_{s,i}\left[\mu R_{s,j} \,\middle|\, Q\right] \leq R_s^{\max}$ (Lemma 3); the equality (47) is due to $\mathbb{P}_s\left[Q=0\right] = 1-\rho$ (Lemma 5). Because $1-\rho \leq 1/n$, (48) implies (44).

Case 2: $\rho < 1 - 1/n$**.** This is the most non-trivial part of the proof. Because $1/n < 1-\rho$, we need to to bound $\Gamma_{s,j}$ by $O(1/n)$ for each $j \in [n]$. To this end, observe that when $\rho < 1 - 1/n$, the load of the j-th leave-one-out system is less than 1, so it has a well-defined stationary distribution (Assumption 3). We consider the indicator $\mathbb{1}\left\{\widetilde{Q}_{-j} = 0\right\}$ as a "baseline" and subtract it from $\mathbb{1}\{Q=0\}$. This leads to the following *key decomposition* of $\mathbb{E}_s\left[\mathbb{1}\{Q=0\}\mu R_{s,j}\right]$:

$$\mathbb{E}_s\left[\mathbb{1}\{Q=0\}\mu R_{s,j}\right] = \mathbb{E}_s\left[\left(\mathbb{1}\{Q=0\} - \mathbb{1}\left\{\widetilde{Q}_{-j}=0\right\}\right)\mu R_{s,j}\right] + \mathbb{E}_s\left[\mathbb{1}\left\{\widetilde{Q}_{-j}=0\right\}\mu R_{s,j}\right]. \tag{49}$$

We then bound the two terms on the right-hand side of (49) separately.

We bound the first term on the right-hand side of (49) using similar arguments as in Case 1, i.e., by applying the conditional expectation bound of $R_{s,j}$ proved in Lemma 3:

$$\begin{aligned}
\mathbb{E}_s\left[\left(\mathbb{1}\{Q=0\} - \mathbb{1}\left\{\widetilde{Q}_{-j}=0\right\}\right)\mu R_{s,j}\right] &\leq \frac{1}{n}\sum_{i:\, i\neq j} \mathbb{E}_{s,i}\left[\mathbb{1}\{Q=0\} - \mathbb{1}\left\{\widetilde{Q}_{-j}=0\right\}\right] R_s^{\max} \qquad (50)\\
&\leq \left(\mathbb{P}_s\left[Q=0\right] - \frac{1}{n}\sum_{i:\, i\neq j}\mathbb{P}_{s,i}\left[\widetilde{Q}_{-j}=0\right]\right) R_s^{\max} \qquad (51)\\
&= \left(1-\rho - \left(1-\rho-\frac{1}{n}\right)\right) R_s^{\max} \qquad (52)\\
&= \frac{1}{n} R_s^{\max},
\end{aligned}$$

where the inequality (50) is due to $R_{s,i} = 0$ with probability 1 under $\mathbb{P}_{s,i}$ (Lemma 2), $\mathbb{E}_{s,i}\left[\mu R_{s,j} \,\middle|\, Q, \widetilde{Q}_{-j}\right] \leq R_s^{\max}$ (Lemma 3), and $\mathbb{1}\{Q=0\} - \mathbb{1}\left\{\widetilde{Q}_{-j}=0\right\} \geq 0$ (Lemma 1); the inequality (51) is because $\mathbb{P}_s[Q=0] \triangleq \sum_{i=1}^n \mathbb{P}_{s,i}[Q=0]/n$; the equality (52) is due to $\mathbb{P}_s[Q=0] = 1-\rho$ and $\sum_{i\neq j}\mathbb{P}_{s,i}\left[\widetilde{Q}_{-j}=0\right]/n = 1-\rho-1/n$ (Lemma 5).

To bound the second term on the right-hand side of (49), recall that $\mathbb{1}\left\{\widetilde{Q}_{-j}=0\right\}$ and $R_{s,j}$ are uncorrelated under the distribution $\mathbb{P}_{s,i}$ for $i\neq j$ (Lemma 4), so we can exactly calculate this term as

$$\begin{aligned}
\mathbb{E}_s\left[\mathbb{1}\left\{\widetilde{Q}_{-j}=0\right\}\mu R_{s,j}\right] &= \frac{1}{n}\sum_{i\,:\,i\neq j}\mathbb{E}_{s,i}\left[\mathbb{1}\left\{\widetilde{Q}_{-j}=0\right\}\mu R_{s,j}\right] \\
&= \frac{1}{n}\sum_{i\,:\,i\neq j}\mathbb{P}_{s,i}\left[\widetilde{Q}_{-j}=0\right]\mathbb{E}_\pi\left[\mu R_{s,j}\right] \qquad (53)\\
&= \left(1-\rho-\frac{1}{n}\right)\mathbb{E}_\pi\left[\mu R_{s,j}\right], \qquad (54)
\end{aligned}$$

where (53) is due to Lemma 4, and (54) is due to Lemma 5.

Substituting the above calculations back to (49), we get

$$\mathbb{E}_s\left[\mathbb{1}\{Q=0\}\mu R_{s,j}\right] \leq \frac{1}{n}R_s^{\max} + \left(1-\rho-\frac{1}{n}\right)\mathbb{E}_\pi\left[\mu R_{s,j}\right].$$

Therefore,

$$\Gamma_{s,j} = \mathbb{E}_s\left[\mathbb{1}\{Q=0\}\mu R_{s,j}\right] - (1-\rho)\mathbb{E}_\pi\left[\mu R_{s,j}\right] \leq \frac{1}{n}\left(R_s^{\max} - \mathbb{E}_\pi\left[\mu R_{s,j}\right]\right).$$

Because $\mathbb{E}_\pi\left[\mu R_{s,j}\right] = \mathbb{E}\left[(\mu S)^2\right]/2$, this finishes the proof of (44) when $\rho < 1-1/n$.

Bounding $-\Gamma_a$. Finally, we prove (45), which uses a similar argument as bounding $\Gamma_{s,j}$ in the case where $\rho \geq 1-1/n$:

$$\begin{aligned}
-\Gamma_a &= (1-\rho)\mathbb{E}_\pi[\Lambda R_a] - \mathbb{E}_s[\mathbb{1}\{Q=0\}\Lambda R_a] \\
&= (1-\rho)\mathbb{E}_\pi[\Lambda R_a] - \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{s,i}[\mathbb{1}\{Q=0\}\Lambda R_a] \\
&\leq (1-\rho)\mathbb{E}_\pi[\Lambda R_a] - (1-\rho)R_a^{\min} \qquad (55)\\
&= (1-\rho)\left(\frac{1}{2}\mathbb{E}\left[(\Lambda A)^2\right] - R_a^{\min}\right).
\end{aligned}$$

where the inequality is due to Lemma 3 and Lemma 5. □

REMARK 4. The most non-trivial part of Lemma 8 is the bound

$$\Gamma_{s,j} \leq \frac{1}{n}\left(R_s^{\max} - \mathbb{E}\left[(\mu S)^2\right]/2\right) \quad \forall \rho < 1, n\in\mathbb{Z}_{\geq 1}. \qquad (56)$$

This bound is significant because it implies the mean queue length bound $\mathbb{E}_\pi[Q] \leq R_s^{\max}/(1-\rho)$ for $M/GI/n$ queues after combined with Lemma 7.

Notably, the proof of (56) exhibits a sharp "phase transition" in the proof techniques at the threshold $\rho = 1-1/n$. For the regime $\rho < 1-1/n$, we use our leave-one-out technique. In the complementary regime where $\rho \geq 1-1/n$, the leave-one-out systems are unstable, but a more conventional approach yields $\Gamma_{s,j} \leq (1-\rho)\left(R_s^{\max} - \mathbb{E}\left[(\mu S)^2\right]/2\right)$, which happens to suffice because $1-\rho \leq 1/n$.

The necessity of two distinct techniques for two disjoint parameter regimes is intriguing. Whether this sharp “phase transition” in the proof techniques at $\rho = 1-1/n$ has a deeper reason remains an interesting open question.

8. Generalization to heterogeneous servers. In the previous sections, we have proved a $1/(1-\rho)$-scaling bound for the steady-state mean queue length of the $GI/GI/n$ queue with homogeneous servers. In this section, we generalize our bound to the $GI/GI/n$ queue with fully heterogeneous servers, where the servers could have mutually distinct service-time distributions.

The organization of this section roughly mirrors the high-level structure of Sections 3, 4, 6, and 7. Specifically, we first set up the $GI/GI/n$ queue and modified $GI/GI/n$ queue with heterogeneous servers in Section 8.1. Then in Section 8.2, we state the assumptions and the main theorem, Theorem 2. In Section 8.3, we establish the technical framework and state the supporting lemmas, which then allows us to prove Theorem 2 in Section 8.4.

8.1. *$GI/GI/n$ queue and modified $GI/GI/n$ queue with heterogeneous servers.* We consider the $GI/GI/n$ queue with heterogeneous servers under the FCFS policy. This queueing system consists of a central queue and n servers, where the service-time distributions could be *server-dependent*. Jobs arrive according to a renewal process. A newly-arrived job joins the end of the queue if all servers are busy, or start being served by one of the idle servers. When a job is in service, it occupies the server for a random amount of time following the service-time distribution of the server; the job leaves the system after completing the service. When a server completes a job, it immediately starts serving a job from the head of the queue if the queue is non-empty. We allow an *arbitrary routing policy* for assigning jobs to idle servers when a new job arrives or when multiple servers free up simultaneously.

The definitions of initial state and sample paths of the $GI/GI/n$ queue can be straightforwardly generalized to the heterogeneous setting, which we do not repeat here. We use A to denote a generic interarrival time, and use S_i to denote a generic service time of the i-th server. We denote the service rate of the i-th server as $\mu_i \triangleq 1/\mathbb{E}[S_i]$, and define the total service rate as $\mu_\Sigma \triangleq \sum_{i=1}^n \mu_i$. Let the load $\rho = \Lambda/\mu_\Sigma$. Let $R_a^{\min} \triangleq \inf_{t\ge 0} \mathbb{E}[\Lambda A - t \mid \Lambda A \ge t]$ and $R_{s,i}^{\max} \triangleq \sup_{t\ge 0} \mathbb{E}[\mu_i S_i - t \mid \mu_i S_i \ge t]$ for $i \in [n]$.

The $GI/GI/n$ queue with heterogeneous servers has the same state representation as the homogeneous setting: at any time $t \ge 0$, its state $X(t)$ is given by

$$X(t) \triangleq (R_{s,1}(t), R_{s,2}(t), \dots R_{s,n}(t), R_a(t), Q(t)),$$

where $R_{s,i}(t)$ is the residual service time of the i-th server, $R_a(t)$ is the residual arrival time, and $Q(t)$ is the queue length. The state space of $\{X(t)\}_{t\ge 0}$ is $\mathcal{X} \triangleq \mathbb{R}^{n+1} \times \mathbb{Z}_{\ge 0}$.

As in the homogeneous setting, here we also consider the modified $GI/GI/n$ queue, which introduces a virtual job whenever a server frees up and the queue is empty. The modified system provides a queue-length upper bound for the original system, as shown in Lemma 9 below. We prove Lemma 9 in Section C.

LEMMA 9. *With the same initial states, at any time t, the queue length of the $GI/GI/n$ queue with heterogeneous server is stochastically dominated by the queue length of the corresponding modified $GI/GI/n$ queue.*

With a slight abuse of notation, we also use $X(t)$ to denote the state of the modified $GI/GI/n$ system. Under the dynamics of the modified $GI/GI/n$ system, $R_{s,i}(t)$ is the forward recurrence time of a renewal process with the inter-event distribution S_i, for each $i \in [n]$; $R_a(t)$ is the forward recurrence time of a renewal process with the inter-event distribution A; the queue length $Q(t)$ is determined by these $(n+1)$ independent renewal processes. The detailed update rule of $X(t)$ is identical to those summarized in Section 3.2.

8.2. *Assumptions and results.* Our results hold under the following three assumptions, which are analogous to Assumptions 1, 2, and 3 in the homogeneous setting. Note that in Assumption 7, the load of a modified $GI/GI/n$ queue refers to the ratio of its arrival rate and the total service rate of its servers.

ASSUMPTION 5. We assume that $\mathbb{E}\left[A^2\right] < \infty$ and $R_{s,i}^{\max} < \infty$ for all $i \in [n]$.

ASSUMPTION 6. For each $i \in [n]$, the service-time distribution S_i is non-lattice.

ASSUMPTION 7. Each modified $GI/GI/n$ queue with heterogeneous servers considered in this section is positive Harris recurrent if its load is less than 1. In addition, let ν be its unique stationary distribution, then $\lim_{T\to\infty} \frac{1}{T}\int_0^T \mathbb{E}\left[Q(t)\right] dt = \mathbb{E}_\nu[Q] < \infty$.

THEOREM 2. *Consider the $GI/GI/n$ queue with heterogeneous servers under the FCFS policy, satisfying Assumptions 5, 6, and 7. When the load $\rho < 1$, let π be the unique stationary distribution of the system. Then the mean queue length under π satisfies*

$$
\begin{aligned}
\mathbb{E}_\pi[Q] \leq{}& \frac{\rho \mathrm{Var}\left[\Lambda A\right] + \sum_{j=1}^n (\mu_j/\mu_\Sigma)\mathrm{Var}\left[\mu_j S_j\right] + 1 - \rho}{2(1-\rho)} \\
&+ \sum_{j=1}^n \min\left(1, \frac{\mu_j}{(1-\rho)\mu_\Sigma}\right)\left(R_{s,j}^{\max} - \frac{1}{2}\mathbb{E}\left[(\mu_j S_j)^2\right]\right) \\
&+ \frac{1}{2}\mathbb{E}\left[(\Lambda A)^2\right] - R_a^{\min}.
\end{aligned} \tag{57}
$$

The bound in (57) implies the following simpler bound after replacing each minimum with $\mu_j/\big((1-\rho)\mu_\Sigma\big)$:

$$
\mathbb{E}_\pi[Q] \leq \frac{\rho \mathrm{Var}\left[\Lambda A\right] - \rho + 2\sum_{j=1}^n (\mu_j/\mu_\Sigma) R_{s,j}^{\max}}{2(1-\rho)} + \frac{1}{2}\mathbb{E}\left[(\Lambda A)^2\right] - R_a^{\min}. \tag{58}
$$

When the arrival process is Poisson, the bound in (58) further simplifies into

$$
\mathbb{E}_\pi[Q] \leq \frac{\sum_{j=1}^n (\mu_j/\mu_\Sigma) R_{s,j}^{\max}}{1-\rho}. \tag{59}
$$

Notably, the bounds in (58) and (59) only depend on the service-time distributions through the *weighted average* of $R_{s,j}^{\max}$ across the servers. This makes these bounds relatively robust, in the sense that having a small fraction of servers with bad tail distributions does not significantly worsen the bounds. In particular, when $\mathrm{Var}\left[\Lambda A\right]$ and $\sum_{j=1}^n (\mu_j/\mu_\Sigma) R_{s,j}^{\max}$ are bounded by constants, these bounds scale as $1/(1-\rho)$ universally for all $\rho < 1$ and n.

8.3. *Technical framework and supporting lemmas.* This subsection has a similar structure as Section 6: we define leave-one-out systems, state the Basic Adjoint Relationship (BAR), and then state four additional supporting lemmas. We also adopt the following no-simultaneous-event assumption to simplify the presentation. The argument for removing this assumption is identical to that provided Remark 2, so we do not repeat it here.

ASSUMPTION 8. In the modified $GI/GI/n$ queue with heterogeneous servers, no two events (i.e., an arrival and a completion, two arrivals, or two completions) happen at the same time.

We first define the leave-one-out systems in the same way as in Section 6.1: recall that for each $j \in [n]$, the j-th leave-one-out system is a modified $GI/GI/n$ queue whose queue length process, $\widetilde{Q}_{-j}(t)$, is driven by the arrival process $\mathbb{R}_a(t)$ and the completion processes $\{R_{s,i}(t)\}_{i\in[n]\setminus\{j\}}$ of the main modified $GI/GI/n$ queue under study.

The following lemma states that the queue length of each leave-one-out system pathwise dominates that of the main system. We omit the proof as it is identical to the one provided for the homogeneous case in Lemma 1.

LEMMA 10. *Consider the modified $GI/GI/n$ queue with heterogeneous servers. For each $j = 1, 2, \ldots n$, the queue length of the j-th leave-one-out system pathwise dominates the queue length of the main system under study, i.e.,*

$$\widetilde{Q}_{-j}(t) \geq Q(t) \quad \forall t \geq 0.$$

Our analysis focuses on the main modified $GI/GI/n$ system coupled with that subset of leave-one-out systems that have well-defined stationary distributions. Specifically, we denote the load of the j-th leave-one-out system as $\rho_{-j} \triangleq \Lambda/(\mu_\Sigma - \mu_j)$ and let $J^s \triangleq \{j \in [n] : \rho_{-j} < 1\}$ be the indices of the leave-one-out systems with subcritical loads. Then we define the state-augmented Markov process, $\overline{X}(t) \triangleq (X(t), \tilde{Q}^s(t))$, where $\tilde{Q}^s = (\widetilde{Q}_{-j})_{j\in J^s}$. The state of $\{\overline{X}(t)\}_{t\geq 0}$ is thus $\overline{\mathcal{X}} \triangleq \mathbb{R}^{n+1}_{\geq 0} \times \mathbb{Z}^{1+|J^s|}_{\geq 0}$.

With the differential operator $\mathcal{A}$ and the Palm distributions defined in the same way as in Section 6.2, we can prove BAR for the heterogeneous setting:

LEMMA 11 (Basic Adjoint Relationship for $GI/GI/n$ with heterogeneous servers). *Consider the modified $GI/GI/n$ system with heterogeneous servers. Suppose Assumptions 5, 6 and 7 hold, and the load $\rho < 1$. For any function $f : \overline{\mathcal{X}} \to \mathbb{R}$ that is bounded, locally Lipschitz continuous, and partially differentiable from the right with respect to its first $(n + 1)$ coordinates, we have*

$$\mathbb{E}_\pi\left[\mathcal{A}f(\overline{X})\right] + \Lambda\mathbb{E}_a\left[f(\overline{X}^+) - f(\overline{X}^-)\right] + \sum_{i=1}^{n} \mu_i \mathbb{E}_{s,i}\left[f(\overline{X}^+) - f(\overline{X}^-)\right] = 0, \tag{60}$$

where $\overline{X}, \overline{X}^+, \overline{X}^- \in \overline{\mathcal{X}}$ are random elements following the distributions specified in the expectation subscripts. Moreover, the Palm distributions $\mathbb{P}_a$ and $\mathbb{P}_{s,i}$ satisfy the following with probability 1:

(i) *If $(\overline{X}^+, \overline{X}^-) \sim \mathbb{P}_a$, then $R_a^- = 0$ and the coordinates of $\overline{X}^+$ and $\overline{X}^-$ satisfy the relations:*

$$R_a^+ = R_a^- + A \tag{61}$$

$$Q^+ = Q^- + 1 \tag{62}$$

$$\widetilde{Q}^+_{-j} = \widetilde{Q}^-_{-j} + 1 \quad \forall j \in J^s, \tag{63}$$

where A represents a generic interarrival time and is independent of $\overline{X}^-$; $J^s \triangleq \{j \in [n] : \rho_{-j} < 1\}$; other coordinates of $\overline{X}^+$ are identical to $\overline{X}^-$.

(ii) *For each $i \in [n]$, if $(\overline{X}^+, \overline{X}^-) \sim \mathbb{P}_{s,i}$, then $R_{s,i}^- = 0$ and the coordinates of $\overline{X}^+$ and $\overline{X}^-$ satisfy the relations:*

$$R_{s,i}^+ = R_{s,i}^- + S_i \tag{64}$$

$$Q^+ = Q^- - \mathbb{1}\{Q^- > 0\} \tag{65}$$

$$\widetilde{Q}^+_{-j} = \widetilde{Q}^-_{-j} - \mathbb{1}\left\{\widetilde{Q}^-_{-j} > 0\right\} \quad \forall j \in J^s\setminus\{i\}, \tag{66}$$

where S_i represents a generic service time of the i-th server and is independent of $\overline{X}^-$; other coordinates of $\overline{X}^+$ are identical to $\overline{X}^-$.

Lemma 11 is proved in Section D.2. To simplify the notation, sometimes we would equivalently write the main equation of BAR (60) as

$$\mathbb{E}_\pi\left[\mathcal{A}f(\overline{X})\right] + \Lambda\mathbb{E}_a\left[f(\overline{X}^+) - f(\overline{X}^-)\right] + \mu_\Sigma\mathbb{E}_s\left[f(\overline{X}^+) - f(\overline{X}^-)\right] = 0, \tag{67}$$

where

$$\mathbb{E}_s[\cdot] \triangleq \sum_{i=1}^n \frac{\mu_i}{\mu_\Sigma}\mathbb{E}_{s,i}[\cdot].$$

We omit the superscript $-$ from $\overline{X}^-$ when it does not appear together with $\overline{X}^+$.

Next, we state the four additional supporting lemmas that generalize Lemmas 3, 4, 5, and 6 to the heterogeneous case. Their proofs are provided in Section D.

LEMMA 12 (Bounded mean residual times, heterogeneous). *Consider the modified $GI/GI/n$ queue with heterogeneous servers. Suppose Assumptions 5 and 7 hold and the load $\rho < 1$. Then for any $j \in [n]$,*

$$\mu_j\mathbb{E}_\nu\left[R_{s,j} \,\middle|\, Q, \tilde{Q}^s\right] \le R_{s,j}^{\max} \tag{27}$$

$$\Lambda\mathbb{E}_\nu\left[R_a \,\middle|\, Q, \tilde{Q}^s\right] \ge R_a^{\min}. \tag{28}$$

where $\mathbb{E}_\nu[\cdot]$ can denote either $\mathbb{E}_\pi[\cdot]$ or $\mathbb{E}_{s,i}[\cdot]$ for $i \neq j$.

LEMMA 13 (Uncorrelation lemma, heterogeneous). *Consider the modified $GI/GI/n$ queue with heterogeneous servers. Suppose Assumptions 5, 6, and 7 hold. Then for any $i, j \in [n]$ such that $i \neq j$ and $\rho_{-j} < 1$,*

$$\mathbb{E}_{s,i}\left[\mathbb{1}\left\{\widetilde{Q}_{-j} = 0\right\}R_{s,j}\right] = \mathbb{P}_{s,i}\left[\widetilde{Q}_{-j} = 0\right]\mathbb{E}_\pi\left[R_{s,j}\right]. \tag{68}$$

LEMMA 14. *Consider the modified $GI/GI/n$ queue with heterogeneous servers. Suppose Assumptions 5 and 7 hold, and the load $\rho < 1$. Then for any $i, j \in [n]$ such that $i \neq j$,*

$$\mathbb{E}_\pi\left[R_{s,i}\right] = \frac{1}{2}\mu_i\mathbb{E}\left[S_i^2\right] \tag{69}$$

$$\mathbb{E}_\pi\left[R_a\right] = \frac{1}{2}\Lambda\mathbb{E}\left[A^2\right] \tag{70}$$

$$\mathbb{E}_{s,i}\left[R_{s,j}\right] + \mathbb{E}_{s,j}\left[R_{s,i}\right] = \frac{1}{2}\mu_i\mathbb{E}\left[S_i^2\right] + \frac{1}{2}\mu_j\mathbb{E}\left[S_j^2\right] \tag{71}$$

$$\mathbb{E}_{s,i}\left[R_a\right] + \mathbb{E}_a\left[R_{s,i}\right] = \frac{1}{2}\mu_{s,i}\mathbb{E}\left[S_i^2\right] + \frac{1}{2}\Lambda\mathbb{E}\left[A^2\right] \tag{72}$$

$$\mathbb{P}_s\left[Q = 0\right] = 1 - \rho \tag{73}$$

$$\frac{1}{\mu_\Sigma}\sum_{i:\, i\neq j}\mu_i\mathbb{P}_{s,i}\left[\widetilde{Q}_{-j} = 0\right] = 1 - \rho - \frac{\mu_j}{\mu_\Sigma} \quad \textit{if } \rho_{-j} < 1. \tag{74}$$

LEMMA 15. *Consider the modified $GI/GI/n$ system with heterogeneous servers. Suppose Assumptions 5 and 7 hold, and the load $\rho < 1$. Then we have $\mathbb{E}_a[Q] < \infty$ and $\mathbb{E}_{s,i}[Q] < \infty$ for all $i \in [n]$.*

8.4. *Proof of Theorem 2.* We are now ready to prove Theorem 2. The proof follows the same two-step structure used for the homogeneous case (Section 7), with the changes being primarily notational. We first derive an expression of $\mathbb{E}_\pi[Q]$ involving covariance terms (Lemma 16), and then bound these terms (Lemma 17) to establish the theorem. To avoid repeating the detailed arguments from the previous section, we present only the key steps of the proof and omit explicit citations to some of the corresponding lemmas.

LEMMA 16. *Consider the modified $GI/GI/n$ system satisfying Assumptions 5 and 7, with the load $\rho < 1$. We have*

$$(1-\rho)\mathbb{E}_\pi[Q] = \frac{\rho \mathrm{Var}[\Lambda A] + \sum_{i=1}^n (\mu_i/\mu_\Sigma)\mathrm{Var}[\mu_i S_i] + 1 - \rho}{2} + \sum_{j=1}^n \Gamma_{s,j} - \Gamma_a\,, \tag{75}$$

where $\Gamma_{s,j}$ and Γ_a are defined as

$$\Gamma_{s,j} = \mathbb{E}_s\left[\mathbb{1}\{Q=0\}\mu_j R_{s,j}\right] - (1-\rho)\mathbb{E}_\pi\left[\mu_j R_{s,j}\right] \quad \forall j \in [n] \tag{76}$$

$$\Gamma_a = \mathbb{E}_s[\mathbb{1}\{Q=0\}\Lambda R_a] - (1-\rho)\mathbb{E}_\pi[\Lambda R_a]. \tag{77}$$

PROOF. For any integer $\kappa > 0$, we take the test function

$$f(\overline{X}) = \Big(Q \wedge \kappa + \sum_{j=1}^n \mu_j R_{s,j} \wedge \kappa - \Lambda R_a \wedge \kappa\Big)^2.$$

in the main equation of BAR (67). After rearranging the terms and taking $\kappa \to \infty$, we get

$$2(\mu_\Sigma - \Lambda)\mathbb{E}_\pi\left[Q + \sum_{j=1}^n \mu_j R_{s,j} - \Lambda R_a\right] \tag{78}$$

$$= \Lambda \mathbb{E}_a\left[2\Big(Q + \sum_{j=1}^n \mu_j R_{s,j}\Big)\Big(1 - \Lambda A\Big) + \Big(1-\Lambda A\Big)^2\right] \tag{79}$$

$$+ \sum_{i=1}^n \mu_i \mathbb{E}_{s,i}\left[2\Big(Q + \sum_{j=1}^n \mu_j R_{s,j} - \Lambda R_a\Big)\Big(\mu_i S_i - \mathbb{1}\{Q>0\}\Big) + \Big(\mu_i S_i \ - \mathbb{1}\{Q>0\}\Big)^2\right]. \tag{80}$$

The interchange of the limit and the expectation is justified by an argument similar to that in the proof of Lemma 7 (see Section 7.1). The key non-trivial condition for this step is the finiteness of the moments $\mathbb{E}_\pi[Q]$, $\mathbb{E}_a[Q]$, and $\mathbb{E}_{s,i}[Q]$, which is guaranteed by Assumption 7 and Lemma 15.

We now simplify the previous equality. First, by direct calculations and applying Lemma 14, the arrival term in (79) simplifies into $\Lambda \mathrm{Var}[\Lambda A]$. Second, each term within the summation in (80) simplifies to:

$$2\mu_i \mathbb{E}_{s,i}\left[\Big(\sum_{j=1}^n \mu_j R_{s,j} - \Lambda R_a\Big)\mathbb{1}\{Q=0\}\right] + \mu_i \mathrm{Var}[\mu_i S_i] + \mu_i(1-\rho).$$

Substituting these expressions back into (79) and (80) yields the following simplified form:

$$2(\mu_\Sigma - \Lambda)\mathbb{E}_\pi\left[Q + \sum_{j=1}^n \mu_j R_{s,j} - \Lambda R_a\right]$$

$$= 2\mu_\Sigma \mathbb{E}_s\left[\Big(\sum_{j=1}^n \mu_j R_{s,j} - \Lambda R_a\Big)\mathbb{1}\{Q=0\}\right] + \Lambda \mathrm{Var}\left[\Lambda A\right] + \sum_{i=1}^n \mu_i \mathrm{Var}\left[\mu_i S_i\right] + \mu_\Sigma(1-\rho).$$

Dividing both sides by $2\mu_\Sigma$ and rearranging the terms complete the proof of Lemma 16. □

LEMMA 17. *Consider the modified $GI/GI/n$ system satisfying Assumptions 5, 6 and 7, with the load $\rho < 1$. Let $\Gamma_{s,j}$ and Γ_a be the terms defined in (37) and (38), respectively. Then*

$$\Gamma_{s,j} \le \min\Big(1-\rho, \frac{\mu_j}{\mu_\Sigma}\Big)\Big(R_{s,j}^{\max} - \frac{1}{2}\mathbb{E}\left[(\mu_j S_j)^2\right]\Big) \tag{81}$$

$$-\Gamma_a \le (1-\rho)\Big(\frac{1}{2}\mathbb{E}\left[(\Lambda A)^2\right] - R_a^{\min}\Big) \tag{82}$$

PROOF. We first bound $\Gamma_{s,j}$ and then $-\Gamma_a$.

Bounding $\Gamma_{s,j}$. We consider two cases based on whether $\rho_{-j} \ge 1$.
Case 1: $\rho_{-j} \ge 1$.

$$\begin{aligned}
\Gamma_{s,j} &= \mathbb{E}_s\left[\mathbb{1}\{Q=0\}\mu_j R_{s,j}\right] - (1-\rho)\mathbb{E}_\pi\left[\mu_j R_{s,j}\right] \\
&= \frac{1}{\mu_\Sigma}\sum_{i=1}^n \mu_i \mathbb{E}_{s,i}\left[\mathbb{1}\{Q=0\}\mu_j R_{s,j}\right] - (1-\rho)\mathbb{E}_\pi\left[\mu R_{s,j}\right] \\
&\le \frac{1}{\mu_\Sigma}\sum_{i\,:\, i\neq j} \mu_i \mathbb{P}_{s,i}\left[Q=0\right] R_{s,j}^{\max} - (1-\rho)\mathbb{E}_\pi\left[\mu R_{s,j}\right] \qquad (83)\\
&\le \mathbb{P}_s\left[Q=0\right] R_{s,j}^{\max} - (1-\rho)\mathbb{E}_\pi\left[\mu R_{s,j}\right] \\
&= (1-\rho) R_{s,j}^{\max} - (1-\rho)\mathbb{E}_\pi\left[\mu R_{s,j}\right] \\
&= (1-\rho)\Big(R_{s,j}^{\max} - \frac{1}{2}\mathbb{E}\left[(\mu_j S_j)^2\right]\Big), \qquad (84)
\end{aligned}$$

where the inequality (83) is by Lemma 12. Because $\Lambda/(\mu_\Sigma - \mu_j) \triangleq \rho_{-j} \ge 1$, we have $1-\rho = 1 - \Lambda/\mu_\Sigma \le \mu_j/\mu_\Sigma$. Consequently, (84) implies (81).
Case 2: $\rho_{-j} < 1$. In this case, the j-th leave-one-out system has a well-defined stationary distribution by Assumption 7. We perform the following decomposition:

$$\begin{aligned}
&\mathbb{E}_s\left[\mathbb{1}\{Q=0\}\mu_j R_{s,j}\right] \\
&\quad = \mathbb{E}_s\left[\Big(\mathbb{1}\{Q=0\} - \mathbb{1}\Big\{\widetilde{Q}_{-j}=0\Big\}\Big)\mu_j R_{s,j}\right] + \mathbb{E}_s\left[\mathbb{1}\Big\{\widetilde{Q}_{-j}=0\Big\}\mu_j R_{s,j}\right].
\end{aligned} \tag{85}$$

We then bound the two terms on the right-hand side of (85) separately.

We bound the first term on the right-hand side of (85) by applying the conditional expectation bound of $R_{s,j}$ in Lemma 12, and note that $\mathbb{1}\{Q=0\} - \mathbb{1}\{\widetilde{Q}_{-j}=0\} \ge 0$ (Lemma 10):

$$\begin{aligned}
\mathbb{E}_s\left[\Big(\mathbb{1}\{Q=0\} - \mathbb{1}\Big\{\widetilde{Q}_{-j}=0\Big\}\Big)\mu_j R_{s,j}\right] &\le \sum_{i\,:\, i\neq j}\frac{\mu_i}{\mu_\Sigma}\mathbb{E}_{s,i}\left[\Big(\mathbb{1}\{Q=0\} - \mathbb{1}\Big\{\widetilde{Q}_{-j}=0\Big\}\Big)\right] R_{s,j}^{\max} \\
&\le \Big(\mathbb{P}_s\left[Q=0\right] - \frac{1}{\mu_\Sigma}\sum_{i\,:\, i\neq j}\mu_i \mathbb{P}_{s,i}\left[\widetilde{Q}_{-j}=0\right]\Big) R_{s,j}^{\max} \\
&= \Big(1-\rho - \Big(1-\rho-\frac{\mu_j}{\mu_\Sigma}\Big)\Big) R_{s,j}^{\max} \\
&= \frac{\mu_j}{\mu_\Sigma} R_{s,j}^{\max}.
\end{aligned}$$

For the second term on the right of (85), recall that $\mathbb{1}\{\widetilde{Q}_{-j}=0\}$ and $R_{s,j}$ are uncorrelated under the distribution $\mathbb{P}_{s,i}$ for $i \neq j$ (Lemma 13), so we can exactly calculate this term as

$$\begin{aligned}\mathbb{E}_s\left[\mathbb{1}\left\{\widetilde{Q}_{-j}=0\right\}\mu_j R_{s,j}\right] &= \sum_{i\,:\, i\neq j}\frac{\mu_i}{\mu_\Sigma}\mathbb{E}_{s,i}\left[\mathbb{1}\left\{\widetilde{Q}_{-j}=0\right\}\mu_j R_{s,j}\right] \\ &= \sum_{i\,:\, i\neq j}\frac{\mu_i}{\mu_\Sigma}\mathbb{P}_{s,i}\left[\widetilde{Q}_{-j}=0\right]\mathbb{E}_\pi\left[\mu_j R_{s,j}\right] \\ &= \left(1-\rho-\frac{\mu_j}{\mu_\Sigma}\right)\mathbb{E}_\pi\left[\mu_j R_{s,j}\right].\end{aligned}$$

Substituting the above calculations back to (85), we get

$$\mathbb{E}_s\left[\mathbb{1}\{Q=0\}\mu_j R_{s,j}\right] \le \frac{\mu_j}{\mu_\Sigma}R_{s,j}^{\max} + \left(1-\rho-\frac{\mu_j}{\mu_\Sigma}\right)\mathbb{E}_\pi\left[\mu R_{s,j}\right].$$

Therefore,

$$\Gamma_{s,j} = \mathbb{E}_s\left[\mathbb{1}\{Q=0\}\mu_j R_{s,j}\right] - (1-\rho)\mathbb{E}_\pi\left[\mu_j R_{s,j}\right] \le \frac{\mu_j}{\mu_\Sigma}\left(R_{s,j}^{\max} - \mathbb{E}_\pi\left[\mu_j R_{s,j}\right]\right).$$

Because $\mathbb{E}_\pi\left[\mu_j R_{s,j}\right] = \mathbb{E}\left[(\mu_j S_j)^2\right]/2$, this finishes the proof of (81) when $\rho_{-j} < 1$.

Bounding $-\Gamma_a$. Finally, we prove (82):

$$\begin{aligned}-\Gamma_a &= (1-\rho)\mathbb{E}_\pi\left[\Lambda R_a\right] - \mathbb{E}_s\left[\mathbb{1}\{Q=0\}\Lambda R_a\right] \\ &= (1-\rho)\mathbb{E}_\pi\left[\Lambda R_a\right] - \sum_{i=1}^n \frac{\mu_i}{\mu_\Sigma}\mathbb{E}_{s,i}\left[\mathbb{1}\{Q=0\}\Lambda R_a\right] \\ &\le (1-\rho)\mathbb{E}_\pi\left[\Lambda R_a\right] - (1-\rho)R_a^{\min} \qquad (86) \\ &= (1-\rho)\left(\frac{1}{2}\mathbb{E}\left[(\Lambda A)^2\right] - R_a^{\min}\right). \qquad (87)\end{aligned}$$

where the inequality is due to the conditional expectation lower bound of R_a in Lemma 12 and the fact that $\mathbb{P}_s\left[Q=0\right] = 1-\rho$. □

9. Conclusion. In this paper, we prove a new bound on the steady-state mean queue length for the $GI/GI/n$ queue. Our bound achieves universal $1/(1-\rho)$-scaling and, while requiring slightly stronger assumptions than those in [53], provides significantly smaller constants. Our proof synthesizes three key elements: a modified $GI/GI/n$ queue, the Basic Adjoint Relationship (BAR), and our new leave-one-out technique. We also extend our proof to analyze the $GI/GI/n$ queue with fully heterogeneous servers.

A natural direction for future work is to investigate the tightness of our bounds. While it may be unrealistic to expect any simple closed-form bound to be arbitrarily tight for the $GI/GI/n$ queue, there is potential for improvement. An ideal bound might resemble Kingman's bound, depending only on the low-order moments of the service-time distribution. Such a bound would yield smaller constants for heavy-tailed systems. Furthermore, in sub-Halfin-Whitt regimes, research in special cases suggests that scalings better than $O(1/(1-\rho))$ may be possible [e.g., 8, 67].

Another promising direction is to apply our proof techniques to other large-scale queueing systems, such as load-balancing models with general service-time distributions.

Finally, it would be interesting to explore the theoretical connections between the leave-one-out technique and other existing techniques for proving asymptotic independence [e.g. 76]. Clarifying their similarities and respective scopes would enrich the analytical toolbox for queueing systems.

REFERENCES

[1] AGHAJANI, R. and RAMANAN, K. (2020). The limit of stationary distributions of many-server queues in the Halfin–Whitt regime. *Math. Oper. Res.* **45** 1016-1055.

[2] ASMUSSEN, S. (2003). *Applied Probability and Queues*. Springer-Verlag New York.

[3] ATAR, R. (2012). A diffusion regime with nondegenerate slowdown. *Oper. Res.* **60** 490-500.

[4] ATAR, R. and SOLOMON, N. (2011). Asymptotically optimal interruptible service policies for scheduling jobs in a diffusion regime with nondegenerate slowdown. *Queueing Syst.* **69** 217–235.

[5] BOROVKOV, A. A. (1965). Some limit theorems in the theory of mass service, II multiple channels systems. *Theory Probab. Appl.* **10** 375-400.

[6] BORST, S., MANDELBAUM, A. and REIMAN, M. I. (2004). Dimensioning large call centers. *Oper. Res.* **52** 17-34.

[7] BRAVERMAN, A. (2017). Stein's method for steady-state diffusion approximations. *arXiv:1704.08398 [math.PR]*.

[8] BRAVERMAN, A. and DAI, J. G. (2016). High order steady-state diffusion approximation of the Erlang-C system. *arxiv:1602.02866 [math.PR]*.

[9] BRAVERMAN, A. and DAI, J. (2017). Stein's method for steady-state diffusion approximations of $M/Ph/n+M$ systems. *Ann. Appl. Probab.* **27** 550-581.

[10] BRAVERMAN, A., DAI, J. G. and FENG, J. (2017). Stein's method for steady-state diffusion approximations: an introduction through the Erlang-A and Erlang-C models. *Stoch. Syst.* **6** 301–366.

[11] BRAVERMAN, A., DAI, J. G. and MIYAZAWA, M. (2017). Heavy traffic approximation for the stationary distribution of a generalized Jackson network: The BAR approach. *Stoch. Syst.* **7** 143–196.

[12] BRAVERMAN, A., DAI, J. and MIYAZAWA, M. (2023). The BAR-approach for multiclass queueing networks with SBP service policies. *arXiv preprint arXiv:2302.05791*.

[13] BRAVERMAN, A., DAI, J. G. and FANG, X. (2024). High-order steady-state diffusion approximations. *Oper. Res.* **72** 604-616.

[14] BRAVERMAN, A., DAI, J. G. and MIYAZAWA, M. (2025). The BAR approach for multiclass queueing networks with SBP service policies. *Stoch. Syst.* **15** 1-49.

[15] BRAVERMAN, A. and SCULLY, Z. (2024). Stein's method and general clocks: diffusion approximation of the $G/G/1$ workload. *arXiv:2407.12716 [math.PR]*.

[16] CHANG, C.-S., THOMAS, J. A. and KIANG, S.-H. (1994). On the stability of open networks: A unified approach by stochastic dominance. *Queueing Syst.* **15** 239–260.

[17] DAI, J. G. (1995). On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Ann. Appl. Probab.* **5** 49 – 77.

[18] DAI, J. G., DIEKER, A. and GAO, X. (2014). Validity of heavy-traffic steady-state approximations in many-server queues with abandonment. *Queueing Syst.* **78** 1–29.

[19] DAI, J. G., GLYNN, P. and XU, Y. (2025). Asymptotic product-form steady-state for generalized Jackson networks in multi-scale heavy traffic. *arXiv:2304.01499 [math.PR]*.

[20] DAI, J. G., GUANG, J. and XU, Y. (2024). Steady-state convergence of the continuous-time JSQ system with general distributions in heavy traffic. *ACM SIGMETRICS Perform. Evaluation Rev.* **52** 39–41.

[21] DAI, J. G., HE, S. and TEZCAN, T. (2010). Many-server diffusion limits for G/Ph/n+GI queues. *Ann. Appl. Probab.* **20** 1854 – 1890.

[22] DAI, J. G. and HUO, D. (2024). Asymptotic product-form steady-state for multiclass queueing networks with SBP service policies in multi-scale heavy traffic. *arXiv:2403.04090 [math.PR]*.

[23] DAI, J. G. and MEYN, S. P. (1995). Stability and convergence of moments for multiclass queueing networks via fluid limit models. *IEEE Trans. Autom. Control* **40** 1889-1904.

[24] DING, L. and CHEN, Y. (2020). Leave-one-out approach for matrix completion: Primal and dual analysis. *IEEE Trans. Inf. Theory* **66** 7274-7301.

[25] DOWNEY, P. J. (1991). Bounding synchronization overhead for parallel iteration. *ORSA J. Comput.* **3** 288-298.

[26] ERYILMAZ, A. and SRIKANT, R. (2012). Asymptotically tight steady-state queue length bounds implied by drift conditions. *Queueing Syst.* **72** 311–359.

[27] GAMARNIK, D. and GOLDBERG, D. A. (2013). Steady-state GI/GI/n queue in the Halfin–Whitt regime. *Ann. Appl. Probab.* **23** 2382–2419.

[28] GAMARNIK, D. and MOMČILOVIĆ, P. (2008). Steady-state analysis of a multiserver queue in the Halfin-Whitt regime. *Adv. Appl. Probab.* **40** 548–577.

[29] GAMARNIK, D. and STOLYAR, A. L. (2012). Multiclass multiserver queueing system in the Halfin–Whitt heavy traffic regime: asymptotics of the stationary distribution. *Queueing Syst.* **71** 25–51.

[30] GARNETT, O., MANDELBAUM, A. and REIMAN, M. (2002). Designing a call center with impatient customers. *Manuf. Serv. Oper. Manag.* **4** 208-227.

[31] GROSOF, I., HARCHOL-BALTER, M. and SCHELLER-WOLF, A. (2022). WCFS: a new framework for analyzing multiserver systems. *Queueing Syst.* **102** 143–174.
[32] GROSOF, I., HONG, Y., HARCHOL-BALTER, M. and SCHELLER-WOLF, A. (2024). The RESET and MARC techniques, with application to multiserver-job analysis. *ACM SIGMETRICS Perform. Evaluation Rev.* **51** 6–7.
[33] GUANG, J., CHEN, X. and DAI, J. G. (2024). Uniform moment bounds for generalized Jackson networks in multiscale heavy traffic. *Math. Oper. Res.* **0** null. Published online ahead of print, DOI: 10.1287/moor.2024.0408.
[34] GURVICH, I., HUANG, J. and MANDELBAUM, A. (2014). Excursion-based universal approximations for the Erlang-A queue in steady-state. *Math. Oper. Res.* **39** 325-373.
[35] HALFIN, S. and WHITT, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29** 567–588.
[36] HOKSTAD, P. (1985). Relations for the workload of the GI/G/s queue. *Advances in Applied Probability* **17** 887–904.
[37] HONG, Y. and SCULLY, Z. (2024). Performance of the Gittins policy in the G/G/1 and G/G/k, with and without setup times. *Perform. Eval.* **163** 102377.
[38] HONG, Y. and WANG, W. (2024). Sharp waiting-time bounds for multiserver jobs. *Stoch. Syst.* **14** 455-478.
[39] IGLEHART, D. L. and WHITT, W. (1970a). Multiple channel queues in heavy traffic. II: sequences, networks, and batches. *Adv. Appl. Probab.* **2** 355–369.
[40] IGLEHART, D. L. and WHITT, W. (1970b). Multiple channel queues in heavy traffic. I. *Adv. Appl. Probab.* **2** 150–177.
[41] JANSSEN, A. J. E. M., VAN LEEUWAARDEN, J. S. H. and ZWART, B. (2008). Corrected asymptotics for a multi-server queue in the Halfin-Whitt regime. *Queueing Syst.* **58** 261–301.
[42] JANSSEN, A. J. E. M., VAN LEEUWAARDEN, J. S. H. and ZWART, B. (2011). Refining square-root safety staffing by expanding Erlang C. *Oper. Res.* **59** 1512-1522.
[43] JELENKOVIĆ, P., MANDELBAUM, A. and MOMČILOVIĆ, P. (2004). Heavy traffic limits for queues with many deterministic servers. *Queueing Syst.* **47** 53–69. 2004,
[44] JIN, X., PANG, G., XU, L. and XU, X. (2025). An approximation to the invariant measure of the limiting diffusion of G/Ph/n + GI queues in the Halfin–Whitt regime and related asymptotics. *Math. Oper. Res.* **50** 783-812.
[45] KELLA, O. and STADJE, W. (2006). Superposition of renewal processes and an application to multi-server queues. *Statistics & Probability Letters* **76** 1914-1924.
[46] KENNEDY, D. P. (1972). Rates of convergence for queues in heavy traffic. II: Sequences of queueing systems. *Adv. Appl. Probab.* **4** 382–391.
[47] KINGMAN, J. F. C. (1962a). Some inequalities for the queue GI/G/1. *Biometrika* **49** 315–324.
[48] KINGMAN, J. F. C. (1962b). On queues in heavy traffic. *J. Roy. Stat. Soc. B Met.* **24** 383–392.
[49] KINGMAN, J. F. C. (1970). Inequalities in the theory of queues. *J. Roy. Stat. Soc. B Met.* **32** 102-110.
[50] KÖLLERSTRÖM, J. (1974). Heavy traffic theory for queues with several servers. I. *J. Appl. Probab.* **11** 544–552.
[51] KÖLLERSTRÖM, J. (1979). Heavy traffic theory for queues with several servers. II. *J. Appl. Probab.* **16** 393–401.
[52] LACHENBRUCH, P. A. and MICKEY, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics* **10** 1–11.
[53] LI, Y. and GOLDBERG, D. A. (2025). Simple and explicit bounds for multiserver queues with $1/(1-\rho)$ scaling. *Math. Oper. Res.* **50** 813-837.
[54] LORDEN, G. (1970). On excess over the boundary. *Ann. Math. Statist.* **41** 520–527.
[55] LOULOU, R. (1973). Multi-channel queues in heavy traffic. *J. Appl. Probab.* **10** 769–777.
[56] MANDELBAUM, A., MASSEY, W. A. and REIMAN, M. I. (1998). Strong approximations for Markovian service networks. *Queueing Syst.* **30** 149–201.
[57] MANDELBAUM, A. and MOMČILOVIĆ, P. (2008a). Queues with many servers: The virtual waiting-time process in the QED regime. *Math. Oper. Res.* **33** 561-586.
[58] MANDELBAUM, A. and MOMČILOVIĆ, P. (2008b). Queues with many servers: The virtual waiting-time process in the qed regime. *Math. Oper. Res.* **33** 561-586.
[59] MEYN, S. P. and TWEEDIE, R. L. (1993). Stability of Markovian processes III: Foster–Lyapunov criteria for continuous-time processes. *Adv. Appl. Probab.* **25** 518–548.
[60] MIYAZAWA, M. (1994a). Decomposition formulas for single server queues with vacations : A unified approach by the rate conservation law. *Commun. Stat. Stoch. Models* **10** 389–413.
[61] MIYAZAWA, M. (1994b). Rate conservation laws: A survey. *Queueing Syst.* **15** 1–58.
[62] MIYAZAWA, M. (2015). Diffusion approximation for stationary analysis of queues and their networks: A review. *J. Operat. Res. Soc. Japan* **58** 104–148.

[63] MORI, M. (1975). Some bounds for queues. *J. Operat. Res. Soc. Japan* **18** 152–181.
[64] NAGAEV, S. V. (1970a). On the speed of convergence in a boundary problem. I. *Theory Probab. Appl.* **15** 163–186.
[65] NAGAEV, S. V. (1970b). On the speed of convergence in a boundary problem. II. *Theory Probab. Appl.* **15** 403–429.
[66] PUHALSKII, A. A. and REIMAN, M. I. (2000). The multiclass GI/PH/N queue in the Halfin-Whitt regime. *Adv. Appl. Probab.* **32** 564–595.
[67] RAJ JHUNJHUNWALA, P., HURTADO-LANGE, D. and THEJA MAGULURI, S. (2024). Exponential tail bounds on queues: A confluence of non-asymptotic heavy traffic and large deviations. *ACM SIGMETRICS Perform. Evaluation Rev.* **51** 18–19.
[68] REED, J. (2009). The G/GI/N queue in the Halfin–Whitt regime. *The Annals of Applied Probability* **19** 2211 – 2269.
[69] SCHELLER-WOLF, A. (2003). Necessary and sufficient conditions for delay moments in FIFO multiserver queues with an application comparing s slow servers with one fast one. *Oper. Res.* **51** 748-758.
[70] SCHELLER-WOLF, A. and SIGMAN, K. (1997). Delay moments for FIFO GI/GI/s queues. *Queueing Syst.* **25** 77–95.
[71] SCHELLER-WOLF, A. and VESILO, R. (2006). Structural interpretation and derivation of necessary and sufficient conditions for delay moments in FIFO multiserver queues. *Queueing Syst.* **54** 221–232.
[72] SCULLY, Z., GROSOF, I. and HARCHOL-BALTER, M. (2020). The Gittins policy is nearly optimal in the M/G/k under extremely general conditions. *Proc. ACM Meas. Anal. Comput. Syst.* **4**.
[73] SHALEV-SHWARTZ, S., SHAMIR, O., SREBRO, N. and SRIDHARAN, K. (2010). Learnability, stability and uniform convergence. *J. Mach. Learn. Res.* **11** 2635–2670.
[74] VESILO, R. and SCHELLER-WOLF, A. (2013). Delay moment bounds for multiserver queues with infinite variance service times. *INFOR* **51** 161–174.
[75] WANG, W., XIE, Q. and HARCHOL-BALTER, M. (2021). Zero queueing for multi-server jobs. *Proc. ACM Meas. Anal. Comput. Syst.* **5**.
[76] WANG, W., HARCHOL-BALTER, M., JIANG, H., SCHELLER-WOLF, A. and SRIKANT, R. (2019). Delay asymptotics and bounds for multitask parallel jobs. *Queueing Syst.* **91** 207–239.
[77] WHITT, W. (1992). Understanding the efficiency of multi-server service systems. *Manage. Sci.* **38** 708-723.
[78] WHITT, W. (2002). *Stochastic-Process Limits*, 2002 ed. *Springer Series in Operations Research and Financial Engineering*. Springer, New York, NY.
[79] WOLFF, R. W. (1987). Upper bounds on work in system for multichannel queues. *J. Appl. Probab.* **24** 547–551.
[80] WORTHINGTON, D. (2009). Reflections on queue modelling from the last 50 years. *J. Oper. Res. Soc.* **60** S83–S92.

APPENDIX A: BOUNDED MEAN RESIDUAL TIME IMPLIES EXPONENTIAL TAIL

In this section, we show that if a distribution has a bounded mean residual time given any age, it must have an exponentially decaying tail, so all of its moments must be finite. This fact has also been proved in [25]. Given this fact, Assumption 1 and Assumption 5 automatically imply that the service-time distributions have exponentially decaying tails and finite moments.

LEMMA 18. *For any non-negative random variable V, suppose there exists a constant $C > 0$ such that*

$$\mathbb{E}[V - t \mid V \geq t] \leq C \quad \forall t \geq 0, \tag{88}$$

Then we have

$$\mathbb{P}[V \geq t] \leq C \exp(-(t-1)/C) \quad \forall t \geq 1, \tag{89}$$

and consequently, $\mathbb{E}[V^m] < \infty$ for all $m \in \mathbb{Z}_{\geq 0}$.

PROOF. Let $\overline{F}(t)$ be the complementary CDF of V. The condition (88) can be rewritten as

$$\int_t^\infty \overline{F}(u)du \leq C\overline{F}(t) \quad \forall t \geq 0. \tag{90}$$

We will mimic the proof of Gronwall's inequality to derive an exponential tail bound for $\overline{F}(t)$. We define the function $G(t)$ as

$$G(t) = \exp(t/C) \int_t^\infty \overline{F}(u)du.$$

Because $\overline{F}(t)$ is bounded on $[0,\infty)$, it is not hard to see that $\int_t^\infty \overline{F}(u)du$ is locally Lipschitz continuous, and thus almost everywhere differentiable. For any t at which $\int_t^\infty \overline{F}(u)du$ is differentiable, $G(t)$ is also differentiable, and $G'(t)$ is given by

$$G'(t) = \exp(t/C)\left(\frac{1}{C}\int_t^\infty \overline{F}(u)du - \overline{F}(t)\right),$$

which is non-positive by (90). Moreover, $G(t)$ is also locally Lipschitz continuous. Therefore, on any compact interval, $G(t)$ is absolutely continuous and satisfies the fundamental theorem of calculus. In particular,

$$G(t) = G(0) + \int_0^t G'(u)du \le G(0) = \int_0^\infty \overline{F}(u)d(u) \le C\overline{F}(0) = C,$$

where the second inequality is due to (90). By the definition of $G(t)$, we have

$$\int_t^\infty \overline{F}(u)du \le C\exp(-t/C).$$

Because $\overline{F}(t)$ is non-increasing and non-negative, we can bound $\overline{F}(t+1)$ as

$$\overline{F}(t+1) \le \int_t^{t+1} \overline{F}(u)du \le C\exp(-t/C).$$

This proves (89). The finiteness of $\mathbb{E}\left[V^m\right]$ follows straightforwardly from the exponential tail bound. □

APPENDIX B: STABILITY OF MODIFIED $GI/GI/n$

In this section, we discuss the conditions under which the modified $GI/GI/n$ queue is positive Harris recurrent and has a finite mean queue length. As mentioned in Remark 1, in addition to $\rho < 1$, some regularity conditions are required on the distributions of the interarrival time and service times. Our results have been proved by directly assuming the positive Harris recurrence and the finiteness of mean queue length (Assumptions 3 and 5). In this section, we will provide a more detailed discussion on these conditions, and sketch why they are likely to imply Assumptions 3 and 5, without giving formal proofs. Throughout this section, we focus on modified $GI/GI/n$ queues with heterogeneous servers and simply refer to them as "modified $GI/GI/n$ queue".

We consider adapting the results of Dai and Meyn [23], which considers queueing networks consisting of single-server stations. In particular, their Theorem 6.2 shows that if such a queueing network is fluid stable and satisfies certain conditions on the interarrival-time and service-time distributions [23, Conditions A1)–A3)], it is is positive Harris recurrence, with certain finite steady-state moments. We conjecture that a similar result can be shown for the modified $GI/GI/n$ queue. Specifically, in addition to Assumption 5, we further assume that the interarrival-time distribution A is *spread out and unbounded*, and the service-time distribution S_i is *spread out* for each server $i \in [n]$. [2] These assumptions are analogous

[2]A distribution ν is spread out if there exists n, such that the distribution of the sum of n independent variables following ν dominates a non-negative absolutely-continuous measure ([2, Section VII.1]).

to Conditions A1)–A3) of [23]. We conjecture that, under these assumptions, a modified $GI/GI/n$ queue is positive Harris recurrent with finite mean queue length whenever $\rho < 1$.

Instead of providing a rigorous full proof, we highlight an intermediate step to provide intuition on why our conjectured result is likely to be true. In the proof of [23, Proposition 5.1], an important fact is is that the queue lengths of the queueing network converge to zero in probability under fluid scaling, i.e.,

$$\lim_{|x|\to\infty} \frac{1}{|x|} |Q^x(t_0 |x|)| = 0 \quad \text{in probability,} \tag{91}$$

where x is a state of the queueing network, $|x|$ represents the sum of all queue lengths and residual times in state x, and $Q^x(t)$ represents the queue length process with the initial state x. This fact follows from the assumed stability of the queueing network's fluid model, utilizing the results in [17, Section IV].

To show (91) for the modified $GI/GI/n$ queue, we invoke the following formula of queue length:

$$Q(t) = \max\left\{Q(0) + N_a(t) - \sum_{i=1}^{n} N_i(t),\ \sup_{0\le t'\le t}\left(N_a(t',t) - \sum_{i=1}^{n} N_i(t',t)\right)\right\}, \tag{92}$$

where $N_a(t)$ denotes the number of real-job arrivals by time t, $N_i(t)$ denotes the number of completions at server i by time t, and $N_a(t',t) \triangleq N_a(t) - N_a(t')$, $N_t(t',t) \triangleq N_i(t) - N_i(t')$. The special case of (92) is proved in [27, Corollary 1], under the additional assumptions that $Q(0) = 0$, the servers are homogeneous and no events occur simultaneously; dropping these assumptions is straightforward from their proof.

We intuitively argue that (92) implies (91): Let $x = (r_1, \dots, r_n, r_a, q)$ be the initial state of the modified $GI/GI/n$ queue. Then (92) implies that

$$\begin{aligned}
&\frac{1}{|x|} Q^x(t_0|x|)\\
&= \max\left\{\frac{1}{|x|}\left(q + N_a(t_0|x|) - \sum_{i=1}^{n} N_i(t_0|x|)\right),\ \frac{1}{|x|}\sup_{0\le t'\le t_0|x|}\left(N_a(t',t_0|x|) - \sum_{i=1}^{n} N_i(t',t_0|x|)\right)\right\}.
\end{aligned} \tag{93}$$

Due to the properties of renewal processes, as $|x|$ gets large, $N_a(t_0|x|) - \sum_{i=1}^{n} N_i(t_0|x|)$ is concentrate around their means $\Lambda t_0|x| - \mu_\Sigma t_0|x|$, if the residual times are initialized at equilibrium. For other initial residual times, with high probability, $N_a(t_0|x|) - \sum_{i=1}^{n} N_i(t_0|x|)$ should not be significantly greater than

$$\Lambda t_0|x| + 1 - \mu_\Sigma\left(t_0|x| - \sum_{i=1}^{n} r_i\right) + o(|x|) = -(\mu_\Sigma - \Lambda)t_0|x| + 1 + \sum_{i=1}^{n}\mu_i r_i + o(|x|).$$

Therefore, intuitively,

$$\begin{aligned}
\frac{1}{|x|}\left(q + N_a(t_0|x|) - \sum_{i=1}^{n} N_i(t_0|x|)\right) &\le \frac{q - (\mu_\Sigma - \Lambda)t_0|x| + 1 + \sum_{i=1}^{n}\mu_i r_i + o(|x|)}{|x|}\\
&\le \max_i \mu_i - (\mu_\Sigma - \Lambda)t_0 + o(1) \quad \text{with high probability,}
\end{aligned}$$

So, taking t_0 to be a sufficiently large constant, the first term inside the maximum of (93) is negative with high probability. Moreover, $\sup_{0\le t'\le t_0|x|}\left(N_a(t',t_0|x|) - \sum_{i=1}^{n} N_i(t',t_0|x|)\right)$ should also scale sublinearly in $|x|$, so the second term inside the maximum of (93) converges to zero in probability as $|x|\to\infty$. We have thus informally justified (91) for the modified $GI/GI/n$ queue.

Intuitively, (91) implies that no matter how large the initial state x is, the queue length becomes significantly shorter after a time proportional to $|x|$. After some technical work that mimic those in Proposition 5.1, 5.3, and 5.4 of [23], one can show that the state of the modified $GI/GI/n$ queue would return to a compact subset of the state space reasonably fast and thus satisfy a certain drift condition (Proposition 6.1 of [23]). Furthermore, the spread out and unbounded conditions of A and S_i's can be used to show that compact subsets under the system's dynamics are petite. Theorem 5.3 of [59] then implies that the modified $GI/GI/n$ queue is positive Harris recurrence, with a finite mean queue length.

APPENDIX C: PROOF OF MODIFIED $GI/GI/n$ QUEUE'S DOMINANCE

In this section, we prove that the queue length distribution of the modified $GI/GI/n$ queue stochastically dominates that of the $GI/GI/n$ queue. We focus on the general case of heterogeneous servers. The proof for homogeneous $GI/GI/n$ can be also found in [27, Proposition 1]; our proof follows essentially the same logic, but we include it for completeness.

LEMMA 9. *With the same initial states, at any time t, the queue length of the $GI/GI/n$ queue with heterogeneous server is stochastically dominated by the queue length of the corresponding modified $GI/GI/n$ queue.*

PROOF. We prove the lemma by constructing a coupling for the sample paths of the $GI/GI/n$ queue and the modified $GI/GI/n$ queue. First, by the assumption of the lemma, two systems have the same initial queue length $Q(0)$, the same set of initially busy servers that are serving real jobs $\mathcal{I}(0)$, and the same initial residual service times for the real jobs in service $(R_{s,i}(0))_{i\in\mathcal{I}(0)}$. For each $k\in\mathbb{Z}_{\geq 1}$, consider the k-th real job that begins service after time 0. We let this job arrive at time τ_a^k in both systems. This job may begin service at different times in the original system and in the modified system, denoted as τ_s^k and $\hat{\tau}_s^k$, respectively. However, when this job begins service, we route it to the same server and let it have the same service time S^k in both systems. Note that here, since we allow the service time distributions to be heterogeneous across servers, the sequence $\{S^k\}_{k=1}^{\infty}$ is not necessarily i.i.d., but this does not complicates our proof.

Let $Q(t)$ and $\hat{Q}(t)$ denote the queue lengths in the original and modified systems, respectively. They can be represented as

$$Q(t)=\left|\{k\in\mathbb{Z}_{\geq 1}\colon \tau_a^k\leq t, \tau_s^k>t\}\right| \tag{94}$$

$$\hat{Q}(t)=\left|\{k\in\mathbb{Z}_{\geq 1}\colon \tau_a^k\leq t, \hat{\tau}_s^k>t\}\right|. \tag{95}$$

Our goal is to show that $Q(t)\leq\hat{Q}(t)$ for all $t\geq 0$. By (94) and (95), it suffices to show that $\tau_s^k\leq\hat{\tau}_s^k$ for all $k\in\mathbb{Z}_{\geq 1}$.

We define some additional notation. For $k\in\mathbb{Z}_{\geq 1}$, we let $Z^k(t)$ be the number of jobs that begins service no later than the k-th job, but has not left the system by time t in the original system. We define $\hat{Z}^k(t)$ similarly for the real jobs in the modified system (excluding the virtual jobs). We let $Z^0(t)$ and $\hat{Z}^0(t)$ be the number of real jobs initially in the two systems and have not left by time t. Formally, for $k\in\mathbb{Z}_{\geq 0}$,

$$Z^k(t)\triangleq\left|\{k'\in\mathbb{Z}_{\geq 1}\colon k'\leq k, \tau_s^{k'}+S^{k'}>t\}\right|+\left|\{i\in\mathcal{I}\colon R_{s,i}(0)>t\}\right|$$

$$\hat{Z}^k(t)\triangleq\left|\{k'\in\mathbb{Z}_{\geq 1}\colon k'\leq k, \hat{\tau}_s^{k'}+S^{k'}>t\}\right|+\left|\{i\in\mathcal{I}\colon R_{s,i}(0)>t\}\right|.$$

We let $\hat{V}(t)$ be the number of virtual jobs in service at time t.

Now we prove by induction that $\tau_s^k\leq\hat{\tau}_s^k$ for all $k\in\mathbb{Z}_{\geq 1}$, i.e., each job in the original $GI/GI/n$ queue begins service no later than in the modified $GI/GI/n$ queue.

- **Base step**: $k=1$. Because job 1 is the first job that begins service after time 0, we have
$$\tau_s^1 = \inf\{t > \tau_a^1 \colon Z^0(t) \le n-1\}$$
$$\hat{\tau}_s^1 = \inf\{t > \tau_a^1 \colon \hat{Z}^0(t) + \hat{V}(t) \le n-1\}.$$
By definitions, $Z^0(t) = \hat{Z}^0(t)$ for all $t \ge 0$, and $\hat{V}(t) \ge 0$, so we have $\tau_s^1 \le \hat{\tau}_s^1$.
- **Induction step**: Suppose we have proved $\tau_s^{k'} \le \hat{\tau}_s^{k'}$ for $1 \le k' \le k$, and we want to show that $\tau_s^{k+1} \le \hat{\tau}_s^{k+1}$. Observe that
$$\tau_s^{k+1} = \inf\{t > \tau_a^{k+1} \colon Z^k(t) \le n-1\}$$
$$\hat{\tau}_s^{k+1} = \inf\{t > \tau_a^{k+1} \colon \hat{Z}^k(t) + \hat{V}(t) \le n-1\}.$$
Because $\tau_s^{k'} + S^{k'} \le \hat{\tau}_s^{k'} + S^{k'}$ for all $1 \le k' \le k$, we have $Z^k(t) \le \hat{Z}^k(t)$ for all $t \ge 0$. Moreover, $\hat{V}(t) \ge 0$, so we get $\tau_s^{k+1} \le \hat{\tau}_s^{k+1}$.

By induction, we have shown that $\tau_s^k \le \hat{\tau}_s^k$ for all $k \in \mathbb{Z}_{\ge 1}$. Combining the induction results and the formula of $Q(t)$ and $\hat{Q}(t)$ in (94) and (95), we have finished the proof. □

APPENDIX D: PROOFS OF SUPPORTING LEMMAS

We prove the supporting lemmas for homogeneous $GI/GI/n$ (Section 6) and heterogeneous $GI/GI/n$ (Section 8) together. This section is organized as follows. We first introduce some additional notation and definitions in Section D.1. Then we prove BAR (Lemmas 2 and 11) in Section D.2. Next, in Section D.3, we prove the lemmas of bounded mean residual times (Lemmas 3 and 12). In Section D.4, we prove the uncorrelation lemmas (Lemmas 4 and 13). Finally, in Section D.5, we prove the preliminary steady-state equalities (Lemmas 5 and 14) and finiteness results (Lemmas 6 and 15).

Throughout Section D, we focus on the modified $GI/GI/n$, whose service-time distributions are allowed to be heterogeneous across servers. We focus on the case where $\rho < 1$, and use the convention that the system is in steady state at time 0, consistent with Section 6.2. Without out loss of generality, we assume that no events (arrivals or completions) happen at the same time point (Assumption 4); see Remark 2 for the necessary changes to proof without this assumption.

D.1. Additional notation and definitions. In this subsection, we introduce some additional notation and definitions regarding the modified $GI/GI/n$ queue and the leave-one-out systems. Since we do not consider the original $GI/GI/n$ queue throughout this section, we will reuse some notation in Section 3 and Section C that was reserved for the original $GI/GI/n$ queue.

Recall that the modified $GI/GI/n$ queue is driven by a set of independent renewal processes corresponding to the arrivals and completions. In the main body, we describe these renewal processes mainly using the residual time processes (i.e., forward recurrence times), $R_a(t)$ and $\{R_{s,i}(t)\}_{i\in[n]}$; here, we define the corresponding ages processes (i.e., backward recurrence times): at each time t,

- Let $\alpha_a(t)$ be the time elapsed since the last arrival.
- Let $\alpha_j(t)$ be the time elapsed since the last completion of the j-th server, for each $j \in [n]$.

Next, we define the alternative state representation based on the ages: we let $\overline{Y}(t) = (Y(t), \tilde{Q}^s(t))$, where $Y(t) \triangleq (\alpha_1(t), \ldots, \alpha_n(t), \alpha_a(t), Q(t))$, and recall that the vector $\tilde{Q}^s$ represents the queue lengths of the set of leave-one-out systems whose loads are below 1.

We let $\overline{Y}(0)$ follow the stationary distribution. Each sample path of $\{\overline{Y}(t)\}_{t\ge 0}$ is determined by the initial state $\overline{Y}(0)$, and the sequences $\{A^k\}_{k\in\mathbb{Z}_{\ge 0}}$ and $\{S_i^k\}_{k\in\mathbb{Z}_{\ge 0}}$ for $i \in [n]$, defined as follows:

- $A^0 - \alpha_a(0)$ is the time of the first arrival, and A^k is the interarrival time between the k-th and $(k+1)$-th arrival events for $k \in \mathbb{Z}_{\geq 1}$. Then $A^k \sim \mathbb{P}[A \in \cdot\,]$ for $k \in \mathbb{Z}_{\geq 1}$, and $A^0 \sim \mathbb{P}[A \in \cdot \mid A \geq \alpha_a(0)]$, where A denotes a generic interarrival time independent of anything else.
- $S_i^0 - \alpha_i(0)$ is the completion time of the initial job on the i-th server, and S_i^k is the time between the k-th and $(k+1)$-th completion events on the i-th server, for each $i \in [n]$ and $k \in \mathbb{Z}_{\geq 1}$. Then $S_i^k \sim \mathbb{P}[S_i \in \cdot\,]$ for $k \in \mathbb{Z}_{\geq 1}$, and $S_i^0 \sim \mathbb{P}[S_i \in \cdot \mid S_i \geq \alpha_i(0)]$, where S_i denotes a generic service time of the i-th server and is independent of anything else.

The above-defined random variables and $\overline{Y}(0)$ are "mutually independent" in the following sense: A^0 is independent of $\overline{Y}(0)$ conditioned on $\alpha_a(0)$; S_i^0 is independent of $\overline{Y}(0)$ conditioned on $\alpha_i(0)$; for each $k \in \mathbb{Z}_{\geq 1}$, A^k is independent of everything else, i.e., $\overline{Y}(0), \{S_i^k\}_{i\in[n],\ell\in\mathbb{Z}_{\geq 0}}$ and $\{A^\ell\}_{\ell\neq k}$; for each $k \in \mathbb{Z}_{\geq 1}$ and $i \in [n]$, S_i^k is independent of everything else, i.e., $\overline{Y}(0), \{S_j^\ell\}_{(j,\ell)\neq(i,k)}$, and $\{A^\ell\}_{\ell\in\mathbb{Z}_{\geq 0}}$.

Now we can write some basic quantities using the stochastic primitives defined above:

For each $k \in \mathbb{Z}_{\geq 1}$, the k-th arrival time is $\tau_a^k = \sum_{\ell=0}^{k-1} A^\ell - \alpha_a(0)$, and we let $\tau_a^0 = -\alpha_a(0)$. For each time $t \geq 0$, the number of arrivals by time t is given by $N_a(t) \triangleq \min\{k \in \mathbb{Z}_{\geq 0} \colon \tau_a^k + A^k > t\}$; the age at time t is $\alpha_a(t) = t - \tau_a^{N_a(t)}$; the residual arrival time at time t is $R_a(t) = A^{N_a(t)} - \alpha_a(t)$.

Similarly, on the i-th server, for each $k \in \mathbb{Z}_{\geq 1}$, the k-th completion time is $\tau_i^k = \sum_{\ell=0}^{k-1} S_i^\ell - \alpha_i(0)$, and we let $\tau_i^0 = -\alpha_i(0)$. For each time $t \geq 0$, the number of completions by time t is $N_i(t) = \min\{k \in \mathbb{Z}_{\geq 0} \colon \tau_i^k + S_i^k > t\}$; the age at time t is $\alpha_i(t) = t - \tau_i^{N_i(t)}$; the residual completion time at time t is $R_{s,i}(t) = S_i^{N_i(t)} - \alpha_i(t)$.

We then define a set of filtrations, corresponding to the arrival process, completion processes, and the whole system:

$$\mathcal{F}_{a,t} = \sigma\Big(\alpha_a(0), A^0, A^1, \dots, A^{N_a(t)-1}, \mathbb{1}\big\{A^{N_a(t)} > t - \tau_a^{N_a(t)}\big\}\Big) \quad \forall t \geq 0$$

$$\mathcal{F}_{i,t} = \sigma\Big(\alpha_i(0), S_i^0, S_i^1, \dots, S_i^{N_i(t)-1}, \mathbb{1}\big\{S_i^{N_i(t)} > t - \tau_i^{N_i(t)}\big\}\Big) \quad \forall i \in [n], t \geq 0$$

$$\mathcal{F}_t = \sigma\Big(\overline{Y}(0), \mathcal{F}_{a,t}, \cup_{i\in[n]}\mathcal{F}_{i,t}\Big) \quad \forall t \geq 0.$$

We denote $\mathcal{F}_{a,\infty} = \cup_{t\geq 0}\mathcal{F}_{a,t}$, $\mathcal{F}_{i,\infty} = \cup_{t\geq 0}\mathcal{F}_{i,t}$ for $i \in [n]$, and $\mathcal{F}_\infty = \cup_{t\geq 0}\mathcal{F}_t$.

It is not hard to see that $N_a(t), \alpha_a(t) \in \mathcal{F}_{a,t}$ for $t \geq 0$, but $R_a(t) \neq \mathcal{F}_{a,t}$. Similarly, $N_i(t), \alpha_i(t) \in \mathcal{F}_{i,t}$ for each $i \in [n]$ and $t \geq 0$, but $R_{s,i}(t) \neq \mathcal{F}_{i,t}$. Furthermore, since $Q(t)$ is determined by the initial state and the arrival and completion events by time t, we have $Q(t) \in \mathcal{F}_t$. Similarly, we have $\widetilde{Q}_{-j}(t) \in \sigma\big(\widetilde{Q}_{-j}(0), \mathcal{F}_{a,t}, \cup_{i\neq j}\mathcal{F}_{i,t}\big)$ for $j \in [n]$ and $t \geq 0$.

D.2. Proof of BAR. In this subsection, we prove BAR for the heterogeneous case Lemma 11, which will imply Lemma 2 for the homogeneous case as a special case. The arguments we use are adapted from Section 6 of [14].

LEMMA 11 (Basic Adjoint Relationship for $GI/GI/n$ with heterogeneous servers). *Consider the modified $GI/GI/n$ system with heterogeneous servers. Suppose Assumptions 5, 6 and 7 hold, and the load $\rho < 1$. For any function $f\colon \overline{\mathbb{X}} \to \mathbb{R}$ that is bounded, locally Lipschitz continuous, and partially differentiable from the right with respect to its first $(n+1)$ coordinates, we have*

$$\mathbb{E}_\pi\big[\mathcal{A}f(\overline{X})\big] + \Lambda\mathbb{E}_a\big[f(\overline{X}^+) - f(\overline{X}^-)\big] + \sum_{i=1}^n \mu_i \mathbb{E}_{s,i}\big[f(\overline{X}^+) - f(\overline{X}^-)\big] = 0, \tag{60}$$

where $\overline{X}, \overline{X}^+, \overline{X}^- \in \overline{\mathcal{X}}$ are random elements following the distributions specified in the expectation subscripts. Moreover, the Palm distributions $\mathbb{P}_a$ and $\mathbb{P}_{s,i}$ satisfy the following with probability 1*:*

(i) *If $(\overline{X}^+, \overline{X}^-) \sim \mathbb{P}_a$, then $R_a^- = 0$ and the coordinates of $\overline{X}^+$ and $\overline{X}^-$ satisfy the relations:*

$$R_a^+ = R_a^- + A \tag{61}$$

$$Q^+ = Q^- + 1 \tag{62}$$

$$\widetilde{Q}_{-j}^+ = \widetilde{Q}_{-j}^- + 1 \quad \forall j \in J^s, \tag{63}$$

where A represents a generic interarrival time and is independent of $\overline{X}^-$; $J^s \triangleq \{j \in [n] : \rho_{-j} < 1\}$; other coordinates of $\overline{X}^+$ are identical to $\overline{X}^-$.

(ii) *For each $i \in [n]$, if $(\overline{X}^+, \overline{X}^-) \sim \mathbb{P}_{s,i}$, then $R_{s,i}^- = 0$ and the coordinates of $\overline{X}^+$ and $\overline{X}^-$ satisfy the relations:*

$$R_{s,i}^+ = R_{s,i}^- + S_i \tag{64}$$

$$Q^+ = Q^- - \mathbb{1}\{Q^- > 0\} \tag{65}$$

$$\widetilde{Q}_{-j}^+ = \widetilde{Q}_{-j}^- - \mathbb{1}\left\{\widetilde{Q}_{-j}^- > 0\right\} \quad \forall j \in J^s \backslash \{i\}, \tag{66}$$

where S_i represents a generic service time of the i-th server and is independent of $\overline{X}^-$; other coordinates of $\overline{X}^+$ are identical to $\overline{X}^-$.

PROOF. We let $\overline{X}(0)$ follow the stationary distribution π at time 0. Let $\tau_1, \tau_2, \ldots$ be the times of all arrivals or completion events after time 0, indexed in temporal order. As assumed throughout our proofs, no events occur simultaneously, and we have sketched what should be changed for the general case in Remark 2.

Proving BAR. In the first part of the proof, we prove the main BAR euqation (60) utilizing the fundamental theorem of calculus. We start by examining the differentiability of $f(\overline{X}(t))$ between two consecutive events, τ^k and τ^{k+1}, for each $k \in \mathbb{Z}_{\geq 1}$. Observe that when $t \in (\tau^k, \tau^{k+1})$, $d\overline{X}_\ell(t)/dt = -1$ for $\ell = 1, 2, \ldots, n+1$, and the other coordinates of $\overline{X}(t)$ remain constant. This implies that $\overline{X}(t)$ is Lipschitz continuous for $t \in [\tau^k, \tau^{k+1}]$. Given the local Lipschitz continuity of $f(\cdot)$ and the compactness of $[\tau^k, \tau^{k+1}]$, we deduce that $f(\overline{X}(t))$ is Lipschitz continuous for $t \in [\tau^k, \tau^{k+1}]$. Since Lipschitz continuity implies absolute continuity, $f(\overline{X}(t))$ is almost everywhere differentiable on $[\tau^k, \tau^{k+1}]$. Then for any t where $f(\overline{X}(t))$ is differentiable, the derivative of $f(\overline{X}(t))$ agrees with the right derivative:

$$\frac{df(\overline{X}(t))}{dt} = -\sum_{\ell=1}^{n+1} \partial_\ell^+ f(\overline{X}(t)) \triangleq \mathcal{A}f(\overline{X}(t)).$$

Moreover, because the fundamental theorem of calculus holds for absolutely continuous functions, we have

$$\begin{aligned} f(\overline{X}(\tau^{k+1} \wedge t-)) - f(\overline{X}(\tau^k)) &= \int_{\tau^k}^{\tau^{k+1} \wedge t} \frac{df(\overline{X}(u))}{du} du \\ &= \int_{\tau^k}^{\tau^{k+1} \wedge t} \mathcal{A}f(\overline{X}(u)) du, \end{aligned} \tag{96}$$

for $1 \le k < N(t) \triangleq \min\{k' : \tau^{k'} > t\}$. Similarly, we can apply the fundamental theorem of calculus to $f(\overline{X}(\tau^1 \wedge t)) - f(\overline{X}(0))$. Summing up all the resulting equations, we get

$$f(\overline{X}(t)) - f(\overline{X}(0)) = \int_0^t \mathcal{A} f(\overline{X}(u))du + \sum_{k=1}^{N(t)-1} \big(f(\overline{X}(\tau^k)) - f(\overline{X}(\tau^k -))\big). \tag{97}$$

Rearranging the terms based on different types of events, we get

$$\begin{aligned} f(\overline{X}(t)) - f(\overline{X}(0)) = \int_0^t \mathcal{A} f(\overline{X}(u))du + \sum_{k=1}^{\infty} \big(f(\overline{X}(\tau_a^k)) - f(\overline{X}(\tau_a^k -))\big)\mathbb{1}\{\tau_a^k \le t\} \\ + \sum_{i=1}^{n}\sum_{k=1}^{\infty} \big(f(\overline{X}(\tau_i^k)) - f(\overline{X}(\tau_i^k -))\big)\mathbb{1}\{\tau_i^k \le t\}. \end{aligned} \tag{98}$$

Now we take the expectations on both sides of (98) to get

$$\begin{aligned} \mathbb{E}\left[f(\overline{X}(t)) - f(\overline{X}(0))\right] = \mathbb{E}\left[\int_0^t \mathcal{A} f(\overline{X}(u))du\right] + \mathbb{E}\left[\sum_{k=1}^{\infty} \big(f(\overline{X}(\tau_a^k)) - f(\overline{X}(\tau_a^k -))\big)\mathbb{1}\{\tau_a^k \le t\}\right] \\ + \sum_{i=1}^{n} \mathbb{E}\left[\sum_{k=1}^{\infty} \big(f(\overline{X}(\tau_i^k)) - f(\overline{X}(\tau_i^k -))\big)\mathbb{1}\{\tau_i^k \le t\}\right]. \end{aligned} \tag{99}$$

Because $\overline{X}(0)$ follows the stationary distribution π, $\overline{X}(0) \overset{d}{=} \overline{X}(u) \sim \pi$ for all $u \ge 0$. This implies that

$$\begin{aligned} \mathbb{E}\left[f(\overline{X}(t)) - f(\overline{X}(0))\right] &= 0 \\ \mathbb{E}\left[\int_0^t \mathcal{A} f(\overline{X}(u))du\right] &= \int_0^t \mathbb{E}\left[\mathcal{A} f(\overline{X}(u))\right] du \\ &= t\,\mathbb{E}_\pi\left[\mathcal{A} f(\overline{X})\right], \end{aligned}$$

where the interchange of $\mathbb{E}[\cdot]$ with $\int_0^t \cdot$ is valid because $\mathcal{A} f(\overline{X}(u)) = -\sum_{\ell=1}^{n+1} \partial_\ell^+ f(\overline{X}(u))$ is bounded for $u \in [0,t]$, due to the local Lipschitz continuity of $f(\cdot)$. Therefore, substituting the above two equations back to (99) and taking $t = 1$, we get

$$\begin{aligned} 0 = \mathbb{E}_\pi\left[\mathcal{A} f(\overline{X})\right] + \mathbb{E}\left[\sum_{k=1}^{\infty} \big(f(\overline{X}(\tau_a^k)) - f(\overline{X}(\tau_a^k -))\big)\mathbb{1}\{\tau_a^k \le 1\}\right] \\ + \sum_{i=1}^{n} \mathbb{E}\left[\sum_{k=1}^{\infty} \big(f(\overline{X}(\tau_i^k)) - f(\overline{X}(\tau_i^k -))\big)\mathbb{1}\{\tau_i^k \le 1\}\right]. \end{aligned} \tag{100}$$

Note that the second and third terms of (100) have similar forms as the Palm expectations: as defined in Section 6.2, we have

$$\begin{aligned} \mathbb{E}_a\left[f(\overline{X}^+) - f(\overline{X}^-)\right] &= \frac{1}{\Lambda}\mathbb{E}\left[\sum_{k=1}^{\infty} \big(f(\overline{X}(\tau_a^k)) - f(\overline{X}(\tau_a^k -))\big)\mathbb{1}\{\tau_a^k \le 1\}\right] \\ \mathbb{E}_{s,i}\left[f(\overline{X}^+) - f(\overline{X}^-)\right] &= \frac{1}{\mu_i}\mathbb{E}\left[\sum_{k=1}^{\infty} \big(f(\overline{X}(\tau_i^k)) - f(\overline{X}(\tau_i^k -))\big)\mathbb{1}\{\tau_i^k \le 1\}\right] \quad \forall i \in [n]. \end{aligned}$$

Substituting the definitions into (100), we get the main equation of BAR:

$$\mathbb{E}_\pi\left[\mathcal{A} f(\overline{X})\right] + \Lambda\mathbb{E}_a\left[f(\overline{X}^+) - f(\overline{X}^-)\right] + \sum_{i=1}^{n} \mu_i \mathbb{E}_{s,i}\left[f(\overline{X}^+) - f(\overline{X}^-)\right] = 0. \tag{60}$$

Characterizing Palm expectations. In the remainder of the proof, we prove Lemma 11 (i) and Lemma 11 (ii), which characterize the joint distribution of $\overline{X}^{+}$ and $\overline{X}^{-}$ under the Palm expectations $\mathbb{P}_a$ and $\mathbb{P}_{s,i}$ for $i \in [n]$.

We start by proving Lemma 11 (i). First, because $R_a(\tau_a^k-) = 0$ for all $k \in \mathbb{Z}_{\geq 1}$, by the definition of $\mathbb{P}_a$, we have

$$\mathbb{P}_a\left[R_a = 0\right] = \frac{1}{\Lambda}\mathbb{E}\left[\sum_{k=1}^{\infty} \mathbb{1}\left\{R_a(\tau_a^k-) = 0, \tau_a^k \leq 1\right\}\right] = 1.$$

Next, to understand the relations between the coordinates of $\overline{X}^{+}$ and $\overline{X}^{-}$ under $\mathbb{P}_a$, we examine the expectation $\mathbb{E}_a\left[f(\overline{X}^{-})g(\overline{X}^{+} - \overline{X}^{-})\right]$ for any bounded measurable functions $f, g \colon \mathbb{R}^{n+1} \times \mathbb{Z}^{1+|J^s|} \to \mathbb{R}$:

$$\begin{aligned}\mathbb{E}_a\left[f(\overline{X}^{-})g(\overline{X}^{+} - \overline{X}^{-})\right] &= \frac{1}{\Lambda}\mathbb{E}\left[\sum_{k=1}^{\infty} f\big(\overline{X}(\tau_a^k-)\big)g\big(\overline{X}(\tau_a^k) - \overline{X}(\tau_a^k-)\big)\mathbb{1}\left\{\tau_a^k \leq 1\right\}\right]\\ &= \frac{1}{\Lambda}\sum_{k=1}^{\infty}\mathbb{E}\left[f\big(\overline{X}(\tau_a^k-)\big)g\big(\overline{X}(\tau_a^k) - \overline{X}(\tau_a^k-)\big)\mathbb{1}\left\{\tau_a^k \leq 1\right\}\right].\end{aligned}$$

By the definitions of the sample paths, for each $k \in \mathbb{Z}_{\geq 1}$, at the time of arrival τ_a^k,

$$\overline{X}(\tau_a^k) = \overline{X}(\tau_a^k-) + A^k \boldsymbol{e}_a + \boldsymbol{e}_q + \sum_{j \in J^s} \boldsymbol{e}_{q,-j}, \tag{101}$$

where $\boldsymbol{e}_a, \boldsymbol{e}_q, \boldsymbol{e}_{q,-j} \in \overline{\mathcal{X}}$ are one-hot vectors with the non-zero entries at the coordinates of R_a, Q, and $\widetilde{Q}_{-j}$, respectively. Note that A^k is independent of $(\tau_a^k, \overline{X}(\tau_a^k-))$ and has the same distribution as A, so

$$\begin{aligned}\mathbb{E}_a\left[f(\overline{X}^{-})g(\overline{X}^{+} - \overline{X}^{-})\right] &= \frac{1}{\Lambda}\sum_{k=1}^{\infty}\mathbb{E}\left[f\big(\overline{X}(\tau_a^k-)\big)\mathbb{1}\left\{\tau_a^k \leq 1\right\}\right]\mathbb{E}\left[g\Big(A\boldsymbol{e}_a + \boldsymbol{e}_q + \sum_{j \in J^s}\boldsymbol{e}_{q,-j}\Big)\right]\\ &= \mathbb{E}_a\left[f\big(\overline{X}^{-}\big)\right]\mathbb{E}\left[g\Big(A\boldsymbol{e}_a + \boldsymbol{e}_q + \sum_{j \in J^s}\boldsymbol{e}_{q,-j}\Big)\right]. \end{aligned}\tag{102}$$

Because f and g can be arbitrary bounded measurable functions, (102) implies that under $\mathbb{P}_a$, $\overline{X}^{+} - \overline{X}^{-}$ is independent of $\overline{X}^{-}$, and has the same distribution as $A\boldsymbol{e}_a + \boldsymbol{e}_q + \sum_{j \in J^s}\boldsymbol{e}_{q,-j}$. This proves Lemma 11 (i).

To show Lemma 11 (ii), note that for each $i \in [n]$ and $k \in \mathbb{Z}_{\geq 1}$, we have $R_{s,i}(\tau_i^k-) = 0$, which implies that $R_{s,i}^{-} = 0$ with probability 1 under $\mathbb{P}_{s,i}$. Moreover, by definitions,

$$\overline{X}(\tau_i^k) = \overline{X}(\tau_i^k-) + S_i^k\boldsymbol{e}_i - \mathbb{1}\left\{Q(\tau_i^k-) > 0\right\}\boldsymbol{e}_q - \sum_{j \in J^s\setminus\{i\}}\mathbb{1}\left\{\widetilde{Q}_{-j}(\tau_i^k-) > 0\right\}\boldsymbol{e}_{q,-j}, \tag{103}$$

where $\boldsymbol{e}_i \in \overline{\mathcal{X}}$ denotes the one-hot vector with the non-zero entry at the coordinate of $R_{s,i}$. Because S_i^k is independent of $(\tau_i^k, \overline{X}(\tau_i^k-))$ and has the same distribution as S_i, for any bounded measurable functions $f \colon \mathbb{R}^{n+1} \times \mathbb{Z}^{1+|J^s|} \to \mathbb{R}$ and $g \colon \mathbb{R} \to \mathbb{R}$, we have

$$\begin{aligned}&\mathbb{E}_{s,i}\left[f(\overline{X}^{-})g(R_{s,i}^{+} - R_{s,i}^{-})\right]\\ &\quad= \frac{1}{\Lambda}\sum_{k=1}^{\infty}\mathbb{E}\left[f\big(\overline{X}(\tau_i^k-)\big)g\big(R_{s,i}(\tau_i^k) - R_{s,i}(\tau_i^k-)\big)\mathbb{1}\left\{\tau_i^k \leq 1\right\}\right]\\ &\quad= \frac{1}{\Lambda}\sum_{k=1}^{\infty}\mathbb{E}\left[f\big(\overline{X}(\tau_i^k-)\big)\mathbb{1}\left\{\tau_i^k \leq 1\right\}\right]\mathbb{E}\big[g(S_i)\big]\end{aligned}$$

$$= \mathbb{E}_{s,i}\left[f(\overline{X}^-)\right]\mathbb{E}\left[g(S_i)\right]. \tag{104}$$

This implies that under $\mathbb{P}_{s,i}$, $R^+_{s,i} - R^-_{s,i}$ is independent of $\overline{X}^-$ and has the same distribution as S_i. In addition, for any bounded measurable function $f\colon \mathbb{R}^{n+1}\times\mathbb{Z}^{1+|J^s|}\times\mathbb{Z}\to\mathbb{R}$,

$$\begin{aligned}
\mathbb{E}_{s,i}\left[f(\overline{X}^-, Q^+)\right] &= \frac{1}{\Lambda}\sum_{k=1}^{\infty}\mathbb{E}\left[f(\overline{X}(\tau_i^k-), Q(\tau_i^k))\mathbb{1}\{\tau_i^k \le 1\}\right] \\
&= \frac{1}{\Lambda}\sum_{k=1}^{\infty}\mathbb{E}\left[f\Big(\overline{X}(\tau_i^k-),\, Q(\tau_i^k-) - \mathbb{1}\big\{Q(\tau_i^k-) > 0\big\}\Big)\mathbb{1}\{\tau_i^k \le 1\}\right] \\
&= \mathbb{E}_{s,i}\left[f\big(\overline{X}^-,\, Q^- - \mathbb{1}\{Q^- > 0\}\big)\right],
\end{aligned}$$

which implies that $Q^+ = Q^- - \mathbb{1}\{Q^- > 0\}$ under the distribution $\mathbb{P}_{s,i}$ with probability 1. Similarly, we can show that $\widetilde{Q}^+_{-j} = \widetilde{Q}^-_{-j} - \mathbb{1}\{\widetilde{Q}^-_{-j} > 0\}$ for $j \in J^s\backslash\{i\}$, $R^+_a = R^-_a$, and $R^+_{s,j} = R^-_{s,j}$ for $j \in [n]\backslash\{i\}$ under $\mathbb{P}_{s,i}$ with probability 1. This proves Lemma 11 (ii). □

D.3. Proof of bounded mean residual times (Lemmas 3 and 12). In this section, we prove Lemma 3 and Lemma 12, restated as follows:

LEMMA 3 (Bounded mean residual times). *Consider the modified $GI/GI/n$ queue satisfying Assumptions 1 and 3, with the load $\rho < 1$. If $\rho < 1 - 1/n$, for any $j \in [n]$,*

$$\mu\mathbb{E}_\nu\left[R_{s,j}\,\middle|\,Q,\widetilde{Q}\right] \le R_s^{\max} \tag{27}$$

$$\Lambda\mathbb{E}_\nu\left[R_a\,\middle|\,Q,\widetilde{Q}\right] \ge R_a^{\min} \tag{28}$$

where $\mathbb{E}_\nu[\cdot]$ can denote either $\mathbb{E}_\pi[\cdot]$ or $\mathbb{E}_{s,i}[\cdot]$ for $i \ne j$; if $1 - 1/n \le \rho < 1$, we have analogs of (27) and (28) without conditioned on $\widetilde{Q}$.

LEMMA 12 (Bounded mean residual times, heterogeneous). *Consider the modified $GI/GI/n$ queue with heterogeneous servers. Suppose Assumptions 5 and 7 hold and the load $\rho < 1$. Then for any $j \in [n]$,*

$$\mu_j\mathbb{E}_\nu\left[R_{s,j}\,\middle|\,Q,\tilde{Q}^s\right] \le R_{s,j}^{\max} \tag{27}$$

$$\Lambda\mathbb{E}_\nu\left[R_a\,\middle|\,Q,\tilde{Q}^s\right] \ge R_a^{\min}. \tag{28}$$

where $\mathbb{E}_\nu[\cdot]$ can denote either $\mathbb{E}_\pi[\cdot]$ or $\mathbb{E}_{s,i}[\cdot]$ for $i \ne j$.

Below, we focus on proving Lemma 12 since it is a strict generalization of Lemma 3.

PROOF OF LEMMA 12. We first show (27) for $\nu = \pi$. By the construction of the sample paths in Section D.1 and the tower property of conditional expectations,

$$\begin{aligned}
\mathbb{E}_\pi\left[R_{s,j}\,\middle|\,Q,\tilde{Q}^s\right] &= \mathbb{E}\left[R_{s,j}(0)\,\middle|\,Q(0),\tilde{Q}^s(0)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[R_{s,j}(0)\,\middle|\,Q(0),\tilde{Q}^s(0),\alpha_j(0)\right]\,\middle|\,Q(0),\tilde{Q}^s(0)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[S_j^0 - \alpha_j(0)\,\middle|\,Q(0),\tilde{Q}^s(0),\alpha_j(0)\right]\,\middle|\,Q(0),\tilde{Q}^s(0)\right]. \qquad (105)
\end{aligned}$$

The inner conditional expectation in the last expression can be bounded as follows:

$$\mathbb{E}\left[S_j^0 - \alpha_j(0) \,\middle|\, Q(0), \tilde{Q}^s(0), \alpha_j(0)\right] = \mathbb{E}\left[S_j^0 - \alpha_j(0) \,\middle|\, \alpha_j(0)\right] \tag{106}$$

$$= \mathbb{E}\left[S_j - \alpha_j(0) \,\middle|\, S_j \geq \alpha_j(0)\right] \tag{107}$$

$$\leq R_{s,j}^{\max}/\mu_j. \tag{108}$$

where (106) is because S_j^0 is by definition independent of $Q(0)$ and $\tilde{Q}^s(0)$ given $\alpha_j(0)$; (107) is by the distribution of S_j^0; (108) is due to the definition $R_{s,j}^{\max} \triangleq \sup_{t'\geq 0} \mathbb{E}\left[\mu_i S_i - t' \mid \mu_i S_i \geq t'\right]$. Substituting the bound in (108) back to (105) yields $\mu_j \mathbb{E}_\pi\left[R_{s,j} \,\middle|\, Q, \tilde{Q}^s\right] \leq R_{s,j}^{\max}$.

Next, we prove (27) for $\nu = \mathbb{P}_{s,i}$ with $i \neq j$. Let $(q, \tilde{q}^s)$ be any realization of $(Q, \tilde{Q}^s)$. By the definition of $\mathbb{E}_{s,i}[\cdot]$, we have

$$\begin{aligned}
&\mathbb{E}_{s,i}\left[R_{s,j}\mathbb{1}\left\{(Q,\tilde{Q}^s) = (q,\tilde{q}^s)\right\}\right]\\
&= \frac{1}{\mu_i}\mathbb{E}\left[\sum_{k=1}^{\infty}\mathbb{1}\left\{\tau_i^k \leq 1, \left(Q(\tau_i^k-), \tilde{Q}^s(\tau_i^k-)\right) = (q,\tilde{q}^s)\right\} R_{s,j}(\tau_i^k-)\right]\\
&= \frac{1}{\mu_i}\sum_{k=1}^{\infty}\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left\{\tau_i^k \leq 1, \left(Q(\tau_i^k-), \tilde{Q}^s(\tau_i^k-)\right) = (q,\tilde{q}^s)\right\} R_{s,j}(\tau_i^k-) \,\middle|\, \overline{Y}(0), \mathcal{F}_{a,\infty}, \{\mathcal{F}_{\ell,\infty}\}_{\ell\neq j}, \mathcal{F}_{j,\tau_i^k}\right]\right]
\end{aligned}$$

$$= \frac{1}{\mu_i}\sum_{k=1}^{\infty}\mathbb{E}\left[\mathbb{1}\left\{\tau_i^k \leq 1\right\}\mathbb{1}\left\{\left(Q(\tau_i^k-), \tilde{Q}^s(\tau_i^k-)\right) = (q,\tilde{q}^s)\right\}\mathbb{E}\left[R_{s,j}(\tau_i^k-) \,\middle|\, \overline{Y}(0), \mathcal{F}_{a,\infty}, \{\mathcal{F}_{\ell,\infty}\}_{\ell\neq j}, \mathcal{F}_{j,\tau_i^k}\right]\right], \tag{109}$$

where (109) is because τ_i^k, $Q(\tau_i^k-)$, and $\tilde{Q}^s(\tau_i^k-)$ are measurable with respect to the sigma algebra

$$\sigma\left(\overline{Y}(0), \mathcal{F}_{a,\infty}, \{\mathcal{F}_{\ell,\infty}\}_{\ell\neq j}, \mathcal{F}_{j,\tau_i^k}\right).$$

To bound the conditional expectation of $R_{s,j}(\tau_i^k-)$ in (109), we rewrite it as:

$$\begin{aligned}
&\mathbb{E}\left[R_{s,j}(\tau_i^k-) \,\middle|\, \overline{Y}(0), \mathcal{F}_{a,\infty}, \{\mathcal{F}_{\ell,\infty}\}_{\ell\neq j}, \mathcal{F}_{j,\tau_i^k}\right]\\
&\quad= \mathbb{E}\left[S_j^{N_j(\tau_i^k-)} - \alpha_j(\tau_i^k-) \,\middle|\, \overline{Y}(0), \mathcal{F}_{a,\infty}, \{\mathcal{F}_{\ell,\infty}\}_{\ell\neq j}, \mathcal{F}_{j,\tau_i^k}\right]
\end{aligned}$$

$$\quad= \sum_{m=0}^{\infty}\mathbb{E}\left[\left(S_j^m - \alpha_j(\tau_i^k-)\right)\mathbb{1}\left\{N_j(\tau_i^k-) = m\right\} \,\middle|\, \overline{Y}(0), \mathcal{F}_{a,\infty}, \{\mathcal{F}_{\ell,\infty}\}_{\ell\neq j}, \mathcal{F}_{j,\tau_i^k}\right]. \tag{110}$$

Because τ_i^k is measurable with respect $\sigma\left(\overline{Y}(0), \mathcal{F}_{i,\infty}\right)$, we can rewrite (110) as $\sum_{m=0}^{\infty} U_m(\tau_i^k-)$, where for each $t \geq 0$, $U_m(t)$ is the random variable define as

$$U_m(t) \triangleq \mathbb{E}\left[\left(S_j^m - \alpha_j(t)\right)\mathbb{1}\left\{N_j(t) = m\right\} \,\middle|\, \overline{Y}(0), \mathcal{F}_{a,\infty}, \{\mathcal{F}_{\ell,\infty}\}_{\ell\neq j}, \mathcal{F}_{j,t}\right].$$

For each $t \geq 0$ and $m \in \mathbb{Z}_{\geq 0}$, $U_m(t)$ can be bounded as

$$U_m(t) = \mathbb{E}\left[\left(S_j^m - \alpha_j(t)\right)\mathbb{1}\left\{N_j(t) = m\right\} \,\middle|\, \mathcal{F}_{j,t}\right] \tag{111}$$

$$= \mathbb{E}\left[\left(S_j^m - (t-\tau_j^m)\right)\mathbb{1}\left\{\tau_j^m \leq t < \tau_j^m + S_j^m\right\} \,\middle|\, \alpha_j(0), \{S_j^\ell\}_{\ell=0}^{m-1}, \mathbb{1}\left\{S_j^m > (t-\tau_j^m)\right\}\right] \tag{112}$$

$$= \left(\mathbb{E}\left[S_j^m \,\middle|\, \alpha_j(0), \{S_j^\ell\}_{\ell=0}^{m-1}, \mathbb{1}\{S_j^m > (t-\tau_j^m)\}\right] - (t-\tau_j^m)\right) \cdot \mathbb{1}\{\tau_j^m \le t < \tau_j^m + S_j^m\}$$

$$= \left(\mathbb{E}\left[S_j \,\middle|\, S_j > t-\tau_j^m\right] - (t-\tau_j^m)\right) \cdot \mathbb{1}\{\tau_j^m \le t < \tau_j^m + S_j^m\} \tag{113}$$

$$\le (R_{s,j}^{\max}/\mu_j) \cdot \mathbb{1}\{\tau_j^m \le t < \tau_j^m + S_j^m\}. \tag{114}$$

where (111) is because $(S_j^m, \alpha_j(t), N_j(t))$ are conditionally independent of $\sigma\left(\overline{Y}(0), \mathcal{F}_{a,\infty}, \{\mathcal{F}_{\ell,\infty}\}_{\ell\neq j}\right)$ given $\alpha_j(0)$, and $\alpha_j(0) \in \mathcal{F}_{j,t}$; (112) is due to the definitions of $N_j(t)$ and $\mathcal{F}_{j,t}$; (113) is by the definition of S_j^m and its independence with $\left(\alpha_j(0), \{S_j^\ell\}_{\ell=0}^{m-1}, \tau_j^m\right)$ if $m \ge 1$; (114) follows from the definition $R_{s,j}^{\max} \triangleq \sup_{t'\ge 0} \mathbb{E}\left[\mu_j S_j - t' \,\middle|\, \mu_i S_j \ge t'\right]$. Substituting (114) back to (110) implies that for $k \in \mathbb{Z}_{\ge 1}$,

$$\begin{aligned}
&\mathbb{E}\left[R_{s,j}(\tau_i^k-) \,\middle|\, \overline{Y}(0), \mathcal{F}_{a,\infty}, \{\mathcal{F}_{\ell,\infty}\}_{\ell\neq j}, \mathcal{F}_{j,\tau_i^k}\right] \\
&\quad \le \sum_{m=0}^{\infty} U_m(\tau_i^k-) \\
&\quad \le (R_{s,j}^{\max}/\mu_j) \cdot \sum_{m=0}^{\infty} \mathbb{1}\{\tau_j^m < \tau_i^k \le \tau_j^m + S_j^m\} \\
&\quad = (R_{s,j}^{\max}/\mu_j) \cdot \mathbb{1}\left\{\tau_j^0 < \tau_i^k < \infty\right\} \\
&\quad \le R_{s,j}^{\max}/\mu_j \quad a.s.
\end{aligned}$$

Further substituting the above inequality back to (109) yields

$$\begin{aligned}
&\mathbb{E}_{s,i}\left[R_{s,j}\mathbb{1}\{(Q,\tilde{Q}^s) = (q,\tilde{q}^s)\}\right] \\
&\quad \le (R_{s,j}^{\max}/\mu_j) \cdot \frac{1}{\mu_i} \sum_{k=1}^{\infty} \mathbb{E}\left[\mathbb{1}\{\tau_i^k \le 1\}\mathbb{1}\{\left(Q(\tau_i^k-), \tilde{Q}^s(\tau_i^k-)\right) = (q,\tilde{q}^s)\}\right] \\
&\quad = (R_{s,j}^{\max}/\mu_j) \cdot \mathbb{P}_{s,i}\left[\left(Q,\tilde{Q}^s\right) = (q,\tilde{q}^s)\right],
\end{aligned}$$

where $(q,\tilde{q}^s)$ is an arbitrary realization of $(Q,\tilde{Q}^s)$. This proves $\mathbb{E}_{s,i}\left[R_{s,j} \,\middle|\, Q, \tilde{Q}^s\right] \le R_{s,j}^{\max}/\mu_j$.

The proof for (28) mirrors the proof for (27) with the following two modifications:

1. All quantities associated with the j-th server's completion process are swapped with the corresponding quantities associated with the arrival process.
2. The definition of $R_a^{\min}$ is used in place of $R_{s,j}^{\max}$.

Given the similarity, the detailed proof for (28) is omitted. □

D.4. Proof of the uncorrelation lemmas (Lemmas 4 and 13). In this subsection, we prove the uncorrelation lemmas, Lemma 4 and Lemma 13, restated as follows:

LEMMA 4 (Uncorrelation lemma). *Consider the modified $GI/GI/n$ queue satisfying Assumptions 1, 2, and 3, with the load $\rho < 1 - 1/n$. For any $i, j \in [n]$ such that $i \neq j$, we have*

$$\mathbb{E}_{s,i}\left[\mathbb{1}\left\{\widetilde{Q}_{-j} = 0\right\} R_{s,j}\right] = \mathbb{P}_{s,i}\left[\widetilde{Q}_{-j} = 0\right] \mathbb{E}_\pi\left[R_{s,j}\right]. \tag{29}$$

LEMMA 13 (Uncorrelation lemma, heterogeneous). *Consider the modified $GI/GI/n$ queue with heterogeneous servers. Suppose Assumptions 5, 6, and 7 hold. Then for any $i, j \in [n]$ such that $i \neq j$ and $\rho_{-j} < 1$,*

$$\mathbb{E}_{s,i}\left[\mathbb{1}\left\{\widetilde{Q}_{-j} = 0\right\} R_{s,j}\right] = \mathbb{P}_{s,i}\left[\widetilde{Q}_{-j} = 0\right] \mathbb{E}_\pi\left[R_{s,j}\right]. \tag{68}$$

We focus on the uncorrelation lemma for the heterogeneous setting (Lemma 13), since it is a strict generalization of Lemma 4. The proof of Lemma 13 is based on the following lemma:

LEMMA 19. *Consider the modified $GI/GI/n$ queue with heterogeneous servers. For any $j \in [n]$, consider the residual service time process of the j-th server, $\{R_{s,j}(t)\}_{t\geq 0}$, and assume that it has non-lattice service times. Then for any $\epsilon > 0$, there exists a deterministic $T_0 > 0$ such that for any $t \geq T_0 + R_{s,j}(0)$, we have*

$$\left|\mathbb{E}\left[R_{s,j}(t) \,\middle|\, R_{s,j}(0)\right] - \mathbb{E}_\pi\left[R_{s,j}\right]\right| \leq \epsilon \quad \textit{w.p. } 1.$$

PROOF OF LEMMA 19. We first consider the case when $R_{s,j}(0) = 0$, so $R_{s,j}(t)$ is a pure renewal process. Because the inter-event time S_j of this renewal process is non-lattice (Assumption 6), we can invoke the Key Renewal Theorem [2, Theorem V.4.3]. Restated in our setting, this theorem implies that

$$\lim_{t\to\infty} \mathbb{E}\left[R_{s,j}(t) \,\middle|\, R_{s,j}(0) = 0\right] = \mathbb{E}_\pi\left[R_{s,j}\right],$$

if we can show that $z(t) \triangleq \mathbb{E}\left[\max(S_j - t, 0)\right]$ is *directly Riemann integrable* and satisfies

$$\mathbb{E}\left[R_{s,j}(t) \,\middle|\, R_{s,j}(0) = 0\right] = z(t) + \int_0^t \mathbb{E}\left[R_{s,j}(t-u) \,\middle|\, R_{s,j}(0) = 0\right] F_j(du), \tag{115}$$

where $\int \cdot F_j(du)$ denotes the integral with respect to u when u follows the distribution of S_j.

We first show that $z(t)$ is directly Riemann integrable. By Proposition V.4.1 of [2], $z(t)$ is directly Riemann integrable if it is non-increasing and Lebesgue integrable on $[0, \infty)$. Obviously, $z(t)$ is non-increasing. Moreover, $z(t)$ is non-negative and its Lebesgue integral can be bounded as

$$\begin{aligned}
\int_0^\infty z(t) &= \int_0^\infty \mathbb{E}\left[\max(S_j - t, 0)\right] dt \\
&= \mathbb{E}\left[\int_0^\infty \max(S_j - t, 0) dt\right] \\
&= \frac{1}{2}\mathbb{E}\left[S_j^2\right] \\
&\leq \frac{1}{2}(R_{s,j}^{\max})^2,
\end{aligned}$$

which is finite by Assumption 5 (or Assumption 1 in the homogeneous case).

Next, we show that $z(t)$ satisfies (115). Recall that S_j^1 is the service time of the job that starts service at time $R_{s,j}(0)$. We decompose $\mathbb{E}\left[R_{s,j}(t) \,\middle|\, R_{s,j}(0) = 0\right]$ based on whether $S_j^1 > t$ to get the following:

$$\begin{aligned}
\mathbb{E}\left[R_{s,j}(t) \,\middle|\, R_{s,j}(0) = 0\right] &= \mathbb{E}\left[R_{s,j}(t)\mathbb{1}\left\{S_j^1 > t\right\} \,\middle|\, R_{s,j}(0) = 0\right] + \mathbb{E}\left[R_{s,j}(t)\mathbb{1}\left\{S_j^1 \leq t\right\} \,\middle|\, R_{s,j}(0) = 0\right] \\
&= \mathbb{E}\left[\max(S_j^1 - t, 0)\right] + \int_0^t \mathbb{E}\left[R_{s,j}(t) \,\middle|\, R_{s,j}(0) = 0, S_j^1 = u\right] F_j(du) \\
&= z(t) + \int_0^t \mathbb{E}\left[R_{s,j}(t-u) \,\middle|\, R_{s,j}(0) = 0\right] F_j(du).
\end{aligned}$$

The above arguments verify the assumptions of the Key Renewal Theorem. Therefore, given $R_{s,j}(0) = 0$, for any $\epsilon > 0$, there exists $T_0 > 0$ such that for any $t \geq T_0$,

$$\left|\mathbb{E}\left[R_{s,j}(t) \,\middle|\, R_{s,j}(0) = 0\right] - \mathbb{E}_\pi\left[R_{s,j}\right]\right| \leq \epsilon. \tag{116}$$

In general, when $R_{s,j}(0) \neq 0$, $R_{s,j}(t)$ is a delayed renewal process and we have

$$\mathbb{E}\left[R_{s,j}\big(t + R_{s,j}(0)\big) \,\middle|\, R_{s,j}(0)\right] = \mathbb{E}\left[R_{s,j}(t) \,\middle|\, R_{s,j}(0) = 0\right] \quad \forall t > 0.$$

Substituting the last equation into (116) finishes the proof. □

Now we are ready to prove Lemma 13.

Proof of Lemma 13. By the definition of the Palm expectation $\mathbb{E}_{s,i}[\cdot]$, for arbitrary $T > 0$, we have

$$\begin{aligned}
&\mathbb{E}_{s,i}\left[\mathbb{1}\left\{\widetilde{Q}_{-j} = 0\right\} R_{s,j}\right] \\
&= \frac{1}{\mu_i T}\mathbb{E}\left[\sum_{k=1}^{\infty} \mathbb{1}\left\{\tau_i^k \leq T, \widetilde{Q}_{-j}(\tau_i^k -) = 0\right\} R_{s,j}(\tau_i^k -)\right] \\
&= \frac{1}{\mu_i T}\sum_{k=1}^{\infty} \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left\{\tau_i^k \leq T, \widetilde{Q}_{-j}(\tau_i^k -) = 0\right\} R_{s,j}(\tau_i^k -) \,\middle|\, \overline{Y}(0), \mathcal{F}_{a,\infty}, \{\mathcal{F}_{\ell,\infty}\}_{\ell \neq j}\right]\right]
\end{aligned}$$

(117)

$$= \frac{1}{\mu_i T}\sum_{k=1}^{\infty} \mathbb{E}\left[\mathbb{1}\left\{\tau_i^k \leq T, \widetilde{Q}_{-j}(\tau_i^k -) = 0\right\} \cdot \mathbb{E}\left[R_{s,j}(\tau_i^k -) \,\middle|\, \overline{Y}(0), \mathcal{F}_{a,\infty}, \{\mathcal{F}_{\ell,\infty}\}_{\ell \neq j}\right]\right]$$

(118)

$$= \frac{1}{\mu_i T}\sum_{k=1}^{\infty} \mathbb{E}\left[\mathbb{1}\left\{\tau_i^k \leq T, \widetilde{Q}_{-j}(\tau_i^k -) = 0\right\} \cdot V(\tau_i^k -)\right],$$

where (117) is because τ_i^k and $\widetilde{Q}_{-j}(\tau_i^k -)$ are measurable with respect to $\sigma\Big(\overline{Y}(0), \mathcal{F}_{a,\infty}, \{\mathcal{F}_{\ell,\infty}\}_{\ell \neq j}\Big)$; $V(\cdot)$ is defined as

$$V(t) \triangleq \mathbb{E}\left[R_{s,j}(t) \,\middle|\, \overline{Y}(0), \mathcal{F}_{a,\infty}, \{\mathcal{F}_{\ell,\infty}\}_{\ell \neq j}\right] \quad \forall t \geq 0,$$

and (118) is due to the fact that $\tau_i^k \in \mathcal{F}_{i,\infty}$. Note that because the process $R_{s,j}(t)$ is conditionally independent of $\overline{Y}(0)$ given $\alpha_j(0)$, we can simplify $V(t)$ into

$$V(t) = \mathbb{E}\left[R_{s,j}(t) \,\middle|\, \alpha_j(0)\right]. \tag{119}$$

In later arguments, we will use the fact that

$$V(t) = \mathbb{E}\left[\mathbb{E}\left[R_{s,j}(t) \,\middle|\, R_{s,j}(0)\right] \,\middle|\, \alpha_j(0)\right], \tag{120}$$

which is true because $R_{s,j}(t) = R_{s,j}(0) + \sum_{\ell=1}^{N_j(t)} S_j^\ell - t$ and $\alpha_j(0) = S_j^0 - R_{s,j}(0)$ are conditionally independent given $R_{s,j}(0)$.

Next, we bounds the difference between $V(\tau_i^k -)$ and $\mathbb{E}_\pi\left[R_{s,j}\right]$ for each $k \in \mathbb{Z}_{\geq 1}$. Recall that by Lemma 19, for any $\epsilon > 0$, there exists a deterministic $T_0 > 0$ such that for any $t \geq T_0 + R_{s,j}(0)$, we have

$$\left|\mathbb{E}\left[R_{s,j}(t) \,\middle|\, R_{s,j}(0)\right] - \mathbb{E}_\pi\left[R_{s,j}\right]\right| \leq \epsilon \quad \text{w.p. } 1. \tag{121}$$

Let the random variable $K^\epsilon \triangleq \min\{k' \in \mathbb{Z}_{\geq 1} \colon \tau_i^{k'+1} > T_0 + R_{s,j}(0)\}$. Then by (120) and (121), we have

$$\left|V(\tau_i^k -) - \mathbb{E}_\pi\left[R_{s,j}\right]\right| \leq \epsilon \quad \text{if } k \geq K^\epsilon + 1. \tag{122}$$

Moreover, we argue that

$$\mathbb{E}\left[K^{\epsilon}\right] \leq \mu_i T_0 + \mu_i \mathbb{E}_\pi\left[R_{s,j}\right] + \mathbb{E}\left[(\mu_i S_i)^2\right] < \infty. \tag{123}$$

To see this, recall that $\tau_i^{k'+1} = \sum_{\ell=0}^{k'} S_i^\ell - \alpha_i(0) = \sum_{\ell=1}^{k'} S_i^\ell + R_{s,i}(0)$, where S_i^ℓ are i.i.d. for $\ell \geq 1$. We can thus rewrite $K^\epsilon = \inf\{k' \in \mathbb{Z}_{\geq 1}\colon \sum_{\ell=1}^{k'} S_i^\ell > T_0 + R_{s,j}(0) - R_{s,i}(0)\}$, which shows that K^ϵ is a stopping time with respect to the discrete-time random walk $\left(\sum_{\ell=1}^{k'} S_i^\ell\right)_{k'=1}^\infty$. By Wald's equation and Lorden's inequality [54], we have

$$\begin{aligned}
\mathbb{E}\left[K^\epsilon \,\middle|\, R_{s,j}(0), R_{s,i}(0)\right]\mathbb{E}\left[S_i\right] &= \mathbb{E}\left[\sum_{\ell=1}^{K^\epsilon} S_i^\ell \,\middle|\, R_{s,j}(0), R_{s,i}(0)\right] \\
&\leq \max(T_0 + R_{s,j}(0) - R_{s,i}(0), 0) + \mu_i \mathbb{E}\left[S_i^2\right],
\end{aligned}$$

which implies (123) after rearranging the terms.

Now we can finish the proof utilizing (118), (122), and (123). First, by the definition of $\mathbb{P}_{s,i}\left[\widetilde{Q}_{-j} = 0\right]$, we have that for any $T > 0$,

$$\mathbb{P}_{s,i}\left[\widetilde{Q}_{-j} = 0\right] = \frac{1}{\mu_i T}\mathbb{E}\left[\sum_{k=1}^\infty \mathbb{1}\left\{\tau_i^k \leq T, \widetilde{Q}_{-j}(\tau_i^k-) = 0\right\}\right]$$

Combing the last equality with the equality in (118), we get

$$\begin{aligned}
&\left|\mathbb{E}_{s,i}\left[\mathbb{1}\left\{\widetilde{Q}_{-j} = 0\right\} R_{s,j}\right] - \mathbb{P}_{s,i}\left[\widetilde{Q}_{-j} = 0\right]\mathbb{E}_\pi\left[R_{s,j}\right]\right| \\
&= \frac{1}{\mu_i T}\left|\mathbb{E}\left[\sum_{k=1}^\infty \mathbb{1}\left\{\tau_i^k \leq T, \widetilde{Q}_{-j}(\tau_i^k-) = 0\right\}\left(V(\tau_i^k-) - \mathbb{E}_\pi\left[R_{s,j}\right]\right)\right]\right| \\
&\leq \frac{1}{\mu_i T}\mathbb{E}\left[\sum_{k=1}^\infty \mathbb{1}\left\{\tau_i^k \leq T\right\} \cdot \left|V(\tau_i^k-) - \mathbb{E}_\pi\left[R_{s,j}\right]\right|\right].
\end{aligned}$$

Fixing any $\epsilon > 0$, we break the infinite sum above into two parts based on K^ϵ:

$$\left|\mathbb{E}_{s,i}\left[\mathbb{1}\left\{\widetilde{Q}_{-j} = 0\right\} R_{s,j}\right] - \mathbb{P}_{s,i}\left[\widetilde{Q}_{-j} = 0\right]\mathbb{E}_\pi\left[R_{s,j}\right]\right|$$

$$\leq \frac{1}{\mu_i T}\mathbb{E}\left[\sum_{k=1}^{K^\epsilon} \mathbb{1}\left\{\tau_i^k \leq T\right\}\left|V(\tau_i^k-) - \mathbb{E}_\pi\left[R_{s,j}\right]\right|\right] \tag{124}$$

$$+ \frac{1}{\mu_i T}\mathbb{E}\left[\sum_{k=K^\epsilon+1}^{\infty} \mathbb{1}\left\{\tau_i^k \leq T\right\}\left|V(\tau_i^k-) - \mathbb{E}_\pi\left[R_{s,j}\right]\right|\right] \tag{125}$$

$$\leq \frac{2R_{s,j}^{\max}\mathbb{E}\left[K^\epsilon\right]}{\mu_i T} + \frac{\epsilon}{\mu_i T}\mathbb{E}\left[\sum_{k=K^\epsilon+1}^{\infty} \mathbb{1}\left\{\tau_i^k \leq T\right\}\right] \tag{126}$$

$$\leq \frac{2R_{s,j}^{\max}\mathbb{E}\left[K^\epsilon\right]}{\mu_i T} + \epsilon. \tag{127}$$

Here, to get (126), we bound the expectation in (124) using the argument that

$$\left|V(\tau_i^k-) - \mathbb{E}_\pi\left[R_{s,j}\right]\right| \leq V(\tau_i^k-) + \mathbb{E}_\pi\left[R_{s,j}\right] \leq 2R_{s,j}^{\max} \quad a.s.$$

To get (127), we use the fact that $\mathbb{E}\left[\sum_{k=1}^\infty \mathbb{1}\left\{\tau_i^k \leq T\right\}\right] = \mu_i T$. Now, because $\mathbb{E}\left[K^\epsilon\right] < \infty$, and T and ϵ are arbitrary, first taking $T \to \infty$ and then taking $\epsilon \to 0$ in (127) finish the proof. $\square$

D.5. Proofs of some steady-state equalities and finiteness results. In this subsection, we will prove several basic equalities about the residual times of the completion processes and the idle probability of the queue (Lemmas 5 and 14). We will also prove that modified G/G/n have finite mean queue lengths under the Palm distributions $\mathbb{P}_a$ and $\mathbb{P}_{s,i}$ (Lemmas 6 and 15). All the results in this subsection are derived from BAR (Lemma 11), which we restate below for ease of reference:

LEMMA 11 (Basic Adjoint Relationship for $GI/GI/n$ with heterogeneous servers). *Consider the modified $GI/GI/n$ system with heterogeneous servers. Suppose Assumptions 5, 6 and 7 hold, and the load $\rho < 1$. For any function $f : \overline{\mathcal{X}} \to \mathbb{R}$ that is bounded, locally Lipschitz continuous, and partially differentiable from the right with respect to its first $(n+1)$ coordinates, we have*

$$\mathbb{E}_\pi\left[\mathcal{A}f(\overline{X})\right] + \Lambda\mathbb{E}_a\left[f(\overline{X}^+) - f(\overline{X}^-)\right] + \sum_{i=1}^{n}\mu_i\mathbb{E}_{s,i}\left[f(\overline{X}^+) - f(\overline{X}^-)\right] = 0, \tag{60}$$

where $\overline{X}, \overline{X}^+, \overline{X}^- \in \overline{\mathcal{X}}$ are random elements following the distributions specified in the expectation subscripts. Moreover, the Palm distributions $\mathbb{P}_a$ and $\mathbb{P}_{s,i}$ satisfy the following with probability 1:

(i) *If $(\overline{X}^+, \overline{X}^-) \sim \mathbb{P}_a$, then $R_a^- = 0$ and the coordinates of $\overline{X}^+$ and $\overline{X}^-$ satisfy the relations:*

$$R_a^+ = R_a^- + A \tag{61}$$

$$Q^+ = Q^- + 1 \tag{62}$$

$$\widetilde{Q}_{-j}^+ = \widetilde{Q}_{-j}^- + 1 \quad \forall j \in J^s, \tag{63}$$

where A represents a generic interarrival time and is independent of $\overline{X}^-$; $J^s \triangleq \{j \in [n] : \rho_{-j} < 1\}$; other coordinates of $\overline{X}^+$ are identical to $\overline{X}^-$.

(ii) *For each $i \in [n]$, if $(\overline{X}^+, \overline{X}^-) \sim \mathbb{P}_{s,i}$, then $R_{s,i}^- = 0$ and the coordinates of $\overline{X}^+$ and $\overline{X}^-$ satisfy the relations:*

$$R_{s,i}^+ = R_{s,i}^- + S_i \tag{64}$$

$$Q^+ = Q^- - \mathbb{1}\{Q^- > 0\} \tag{65}$$

$$\widetilde{Q}_{-j}^+ = \widetilde{Q}_{-j}^- - \mathbb{1}\left\{\widetilde{Q}_{-j}^- > 0\right\} \quad \forall j \in J^s \backslash \{i\}, \tag{66}$$

where S_i represents a generic service time of the i-th server and is independent of $\overline{X}^-$; other coordinates of $\overline{X}^+$ are identical to $\overline{X}^-$.

Now, we restate Lemma 5 and its generalization in the heterogeneous setting, Lemma 14. We will directly prove Lemma 14.

LEMMA 5. *Consider the modified $GI/GI/n$ queue satisfying Assumptions 1 and 3, with the load $\rho < 1$. For $i, j \in [n]$ such that $i \neq j$, we have*

$$\mathbb{E}_\pi\left[R_{s,i}\right] = \frac{1}{2}\mu\mathbb{E}\left[S^2\right] \tag{30}$$

$$\mathbb{E}_\pi\left[R_a\right] = \frac{1}{2}\Lambda\mathbb{E}\left[A^2\right] \tag{31}$$

$$\mathbb{E}_{s,i}\left[R_{s,j}\right] + \mathbb{E}_{s,j}\left[R_{s,i}\right] = \mu\mathbb{E}\left[S^2\right] \tag{32}$$

$$\mathbb{E}_{s,i}\left[R_a\right]+\mathbb{E}_a\left[R_{s,i}\right]=\frac{1}{2}\mu\mathbb{E}\left[S^2\right]+\frac{1}{2}\Lambda\mathbb{E}\left[A^2\right] \tag{33}$$

$$\mathbb{P}_s\left[Q=0\right]=1-\rho \tag{34}$$

$$\frac{1}{n}\sum_{i:\,i\neq j}\mathbb{P}_{s,i}\left[\widetilde{Q}_{-j}=0\right]=1-\rho-\frac{1}{n}\quad \textit{if } \rho<1-1/n. \tag{35}$$

LEMMA 14. *Consider the modified $GI/GI/n$ queue with heterogeneous servers. Suppose Assumptions 5 and 7 hold, and the load $\rho<1$. Then for any $i,j\in[n]$ such that $i\neq j$,*

$$\mathbb{E}_\pi\left[R_{s,i}\right]=\frac{1}{2}\mu_i\mathbb{E}\left[S_i^2\right] \tag{69}$$

$$\mathbb{E}_\pi\left[R_a\right]=\frac{1}{2}\Lambda\mathbb{E}\left[A^2\right] \tag{70}$$

$$\mathbb{E}_{s,i}\left[R_{s,j}\right]+\mathbb{E}_{s,j}\left[R_{s,i}\right]=\frac{1}{2}\mu_i\mathbb{E}\left[S_i^2\right]+\frac{1}{2}\mu_j\mathbb{E}\left[S_j^2\right] \tag{71}$$

$$\mathbb{E}_{s,i}\left[R_a\right]+\mathbb{E}_a\left[R_{s,i}\right]=\frac{1}{2}\mu_{s,i}\mathbb{E}\left[S_i^2\right]+\frac{1}{2}\Lambda\mathbb{E}\left[A^2\right] \tag{72}$$

$$\mathbb{P}_s\left[Q=0\right]=1-\rho \tag{73}$$

$$\frac{1}{\mu_\Sigma}\sum_{i:\,i\neq j}\mu_i\mathbb{P}_{s,i}\left[\widetilde{Q}_{-j}=0\right]=1-\rho-\frac{\mu_j}{\mu_\Sigma}\quad \textit{if } \rho_{-j}<1. \tag{74}$$

PROOF OF LEMMA 14. **Proof of (69).** For any truncation threshold $\kappa>0$, consider the test function $f(\overline{X})=(R_{s,i}\wedge\kappa)^2$, which is bounded, locally Lipschitz continuous, and has right partial derivatives. Consequently, Basic Adjoint Relationship (60) implies that

$$-2\mathbb{E}_\pi\left[R_{s,i}\mathbb{1}\left\{R_{s,i}<\kappa\right\}\right]+\Lambda\mathbb{E}_{s,i}\left[\left((R_{s,i}+S_i)\wedge\kappa\right)^2-\left(R_{s,i}\wedge\kappa\right)^2\right]=0.$$

Because $R_{s,i}=0$ with probability 1 under the measure $\mathbb{P}_{s,i}$, rearranging the terms, we get

$$\mathbb{E}_\pi\left[R_{s,i}\mathbb{1}\left\{R_{s,i}<\kappa\right\}\right]=\frac{1}{2}\Lambda\mathbb{E}\left[(S_i\wedge\kappa)^2\right].$$

Observe that $R_{s,i}\mathbb{1}\left\{R_{s,i}<\kappa\right\}$ and $(S_i\wedge\kappa)^2$ are both non-negative and non-decreasing in κ, by the monotone convergence theorem,

$$\mathbb{E}_\pi\left[R_{s,i}\right]=\lim_{\kappa\to\infty}\mathbb{E}_\pi\left[R_{s,i}\mathbb{1}\left\{R_{s,i}<\kappa\right\}\right]=\lim_{\kappa\to\infty}\frac{1}{2}\Lambda\mathbb{E}\left[(S_i\wedge\kappa)^2\right]=\frac{1}{2}\Lambda\mathbb{E}\left[S_i^2\right], \tag{128}$$

which implies (69).

Proof of (70). This proof is identical to the proof of (69) except that $R_{s,i}$ is replaced by R_a.

Proof of (71). For any $\kappa>0$, we take the test function $f(\overline{X})=(R_{s,i}\wedge\kappa)(R_{s,j}\wedge\kappa)$ in (60):

$$\begin{aligned}&-\mathbb{E}_\pi\left[\mathbb{1}\left\{R_{s,i}<\kappa\right\}(R_{s,j}\wedge\kappa)+\mathbb{1}\left\{R_{s,j}<\kappa\right\}(R_{s,i}\wedge\kappa)\right]\\&+\mu_i\mathbb{E}_{s,i}\left[(S_i\wedge\kappa)(R_{s,j}\wedge\kappa)-0\right]+\mu_j\mathbb{E}_{s,j}\left[(S_j\wedge\kappa)(R_{s,i}\wedge\kappa)-0\right]=0,\end{aligned}$$

where in the second and third expectations we have used the facts that $R_{s,\ell}=0$ with probability 1 under $\mathbb{P}_{s,\ell}$ for any $\ell\in[n]$. Because S_i is independent of $R_{s,j}$ under the measures $\mathbb{P}_{s,i}$, and S_j is independent of $R_{s,i}$ under the measure $\mathbb{P}_{s,j}$, we get

$$\begin{aligned}&\mathbb{E}_\pi\left[\mathbb{1}\left\{R_{s,i}<\kappa\right\}(R_{s,j}\wedge\kappa)+\mathbb{1}\left\{R_{s,j}<\kappa\right\}(R_{s,i}\wedge\kappa)\right]\\&\quad=\mu_i\mathbb{E}\left[S_i\wedge\kappa\right]\mathbb{E}_{s,i}\left[R_{s,j}\wedge\kappa\right]+\mu_j\mathbb{E}\left[S_j\wedge\kappa\right]\mathbb{E}_{s,j}\left[R_{s,i}\wedge\kappa\right].\end{aligned}$$

Because $\mathbb{1}\{R_{s,i} < \kappa\}(R_{s,j} \wedge \kappa)$, $\mathbb{1}\{R_{s,j} < \kappa\}(R_{s,i} \wedge \kappa)$, $S_i \wedge \kappa$, $S_j \wedge \kappa$, $R_{s,j} \wedge \kappa$, and $R_{s,i} \wedge \kappa$ are all non-negative and non-decreasing in κ, by the monotone convergence theorem, we have the following after letting $\kappa \to \infty$:

$$\mathbb{E}_\pi\left[R_{s,j} + R_{s,i}\right] = \mu_i \mathbb{E}\left[S_i\right] \mathbb{E}_{s,i}\left[R_{s,j}\right] + \mu_j \mathbb{E}\left[S_j\right] \mathbb{E}_{s,j}\left[R_{s,i}\right].$$

Note that $\mathbb{E}\left[S_i\right] = 1/\mu_i$, $\mathbb{E}\left[S_j\right] = 1/\mu_j$, so the last equality simplifies into

$$\mathbb{E}_\pi\left[R_{s,j} + R_{s,i}\right] = \mathbb{E}_{s,i}\left[R_{s,j}\right] + \mathbb{E}_{s,j}\left[R_{s,i}\right]. \tag{129}$$

By (69), the left-hand side of (129) equals $\mu_i \mathbb{E}[S_i^2]/2 + \mu_j \mathbb{E}[S_j^2]/2$. This proves (71).

Proof of (72). For any $\kappa > 0$, we take the test function $f(\overline{X}) = (R_{s,i} \wedge \kappa)(R_a \wedge \kappa)$ in (60). The rest of the proof is identical to the proof of (71) except that $R_{s,j}$ is replaced by R_a.

Proof of (73). For any integer $\kappa > 0$, taking the test function $f(\overline{X}) = Q \wedge \kappa$ in (60) yields

$$\begin{aligned} \Lambda \mathbb{E}_a[(Q+1) \wedge \kappa - Q \wedge \kappa] + \mu_\Sigma \mathbb{E}_s\left[(Q-1)^+ \wedge \kappa - Q \wedge \kappa\right] &= 0, \\ \implies \quad \Lambda \mathbb{E}_a[\mathbb{1}\{Q \le \kappa - 1\}] - \mu_\Sigma \mathbb{E}_s[\mathbb{1}\{0 < Q \le \kappa\}] &= 0. \end{aligned}$$

Because each term in the expectations is bounded, letting $\kappa \to \infty$ and invoking the bounded convergence theorem, we get

$$\Lambda - \mu_\Sigma \mathbb{P}_s\left[Q > 0\right] = 0. \tag{130}$$

Rearranging the terms yield (34).

Proof of (74). This proof is identical to the proof of (73) except that Q is replaced by $\widetilde{Q}_{-j}$. □

Next, we restate Lemma 6 and its generalization in the heterogeneous setting Lemma 15. We will directly prove Lemma 15.

LEMMA 6. *Consider the modified $GI/GI/n$ system satisfying Assumptions 1 and 3, with the load $\rho < 1$. Then we have $\mathbb{E}_a[Q] < \infty$ and $\mathbb{E}_{s,i}[Q] < \infty$ for all $i \in [n]$.*

LEMMA 15. *Consider the modified $GI/GI/n$ system with heterogeneous servers. Suppose Assumptions 5 and 7 hold, and the load $\rho < 1$. Then we have $\mathbb{E}_a[Q] < \infty$ and $\mathbb{E}_{s,i}[Q] < \infty$ for all $i \in [n]$.*

PROOF OF LEMMA 15. For any positive integer κ, we consider the test function $f(\overline{X}) = (Q \wedge \kappa - 2\Lambda R_a \wedge \kappa)^2$, which is bounded, locally Lipschitz continuous, and has right partial derivatives. Consequently, Basic Adjoint Relationship (60) implies that

$$0 = \mathbb{E}_\pi[2(Q \wedge \kappa - 2\Lambda R_a \wedge \kappa) \cdot (2\Lambda \mathbb{1}\{R_a \le \kappa\})] \tag{131}$$

$$+ \Lambda \mathbb{E}_a\left[\left((Q+1) \wedge \kappa - 2\Lambda A \wedge \kappa\right)^2 - (Q \wedge \kappa)^2\right] \tag{132}$$

$$+ \mu_\Sigma \mathbb{E}_s\left[\left((Q - \mathbb{1}\{Q > 0\}) \wedge \kappa - 2\Lambda R_a \wedge \kappa\right)^2 - (Q \wedge \kappa - 2\Lambda R_a \wedge \kappa)^2\right], \tag{133}$$

where to get the term in (132), we have used the fact that $R_a = 0$ with probability 1 under $\mathbb{P}_a$.

We first simplify the term in (132): substituting $(Q+1) \wedge \kappa = Q \wedge \kappa + \mathbb{1}\{Q \le \kappa - 1\}$ and rearranging the terms, we get

$$\begin{aligned} &\left((Q+1) \wedge \kappa - 2\Lambda A \wedge \kappa\right)^2 - (Q \wedge \kappa)^2 \\ &\quad = 2(Q \wedge \kappa) \cdot (\mathbb{1}\{Q \le \kappa - 1\} - 2\Lambda A \wedge \kappa) + (\mathbb{1}\{Q \le \kappa - 1\} - 2\Lambda A \wedge \kappa)^2. \end{aligned}$$

Because A is independent of Q under the measure $\mathbb{P}_a$, we have

$$\begin{aligned}&\mathbb{E}_a\left[\left((Q+1)\wedge\kappa-2\Lambda A\wedge\kappa\right)^2-(Q\wedge\kappa)^2\right]\\&\quad=2\mathbb{E}_a\left[(Q\wedge\kappa)\mathbb{1}\{Q\le\kappa-1\}\right]-4\mathbb{E}_a\left[Q\wedge\kappa\right]\mathbb{E}\left[\Lambda A\wedge\kappa\right]\\&\qquad+\mathbb{E}_a\left[\left(\mathbb{1}\{Q\le\kappa-1\}-2\Lambda A\wedge\kappa\right)^2\right].\end{aligned}$$

We then simplify the term in (133). Substituting $(Q-\mathbb{1}\{Q>0\})\wedge\kappa=Q\wedge\kappa-\mathbb{1}\{0<Q\le\kappa\}$ and rearranging the terms, we get

$$\begin{aligned}&\mathbb{E}_s\left[\left((Q-\mathbb{1}\{Q>0\})\wedge\kappa-2\Lambda R_a\wedge\kappa\right)^2-(Q\wedge\kappa-2\Lambda R_a\wedge\kappa)^2\right]\\&\quad=\mathbb{E}_s\left[-2(Q\wedge\kappa-2\Lambda R_a\wedge\kappa)\mathbb{1}\{0<Q\le\kappa\}+\mathbb{1}\{0<Q\le\kappa\}\right]\\&\quad=-2\mathbb{E}_s\left[(Q\wedge\kappa)\mathbb{1}\{0<Q\le\kappa\}\right]+4\mathbb{E}_s\left[(\Lambda R_a\wedge\kappa)\mathbb{1}\{0<Q\le\kappa\}\right]+\mathbb{P}_s\left[0<Q\le\kappa\right].\end{aligned}$$

Substituting the above calculations back into (132) and (133), we get

$$\begin{aligned}0&=\mathbb{E}_\pi\left[2(Q\wedge\kappa-2\Lambda R_a\wedge\kappa)\cdot(2\Lambda\mathbb{1}\{R_a\le\kappa\})\right]\\&\quad+2\Lambda\mathbb{E}_a\left[(Q\wedge\kappa)\mathbb{1}\{Q\le\kappa-1\}\right]-4\Lambda\mathbb{E}_a\left[Q\wedge\kappa\right]\mathbb{E}\left[\Lambda A\wedge\kappa\right]+\Lambda\mathbb{E}_a\left[\left(\mathbb{1}\{Q\le\kappa-1\}-2\Lambda A\wedge\kappa\right)^2\right]\\&\quad-2\mu_\Sigma\mathbb{E}_s\left[(Q\wedge\kappa)\mathbb{1}\{0<Q\le\kappa\}\right]+4\mu_\Sigma\mathbb{E}_s\left[(\Lambda R_a\wedge\kappa)\mathbb{1}\{0<Q\le\kappa\}\right]+\mu_\Sigma\mathbb{P}_s\left[0<Q\le\kappa\right].\end{aligned}$$

Moving the terms involving $\mathbb{E}_a[Q\wedge\kappa]$ and $\mathbb{E}_s[Q\wedge\kappa]$ to the left-hand side, we get

$$\begin{aligned}&-2\Lambda\mathbb{E}_a\left[(Q\wedge\kappa)\mathbb{1}\{Q\le\kappa-1\}\right]+4\Lambda\mathbb{E}_a\left[Q\wedge\kappa\right]\mathbb{E}\left[\Lambda A\wedge\kappa\right]+2\mu_\Sigma\mathbb{E}_s\left[(Q\wedge\kappa)\mathbb{1}\{0<Q\le\kappa\}\right]\\&=\mathbb{E}_\pi\left[2(Q\wedge\kappa-2\Lambda R_a\wedge\kappa)\cdot(2\Lambda\mathbb{1}\{R_a\le\kappa\})\right]+\Lambda\mathbb{E}_a\left[\left(\mathbb{1}\{Q\le\kappa-1\}-2\Lambda A\wedge\kappa\right)^2\right]\\&\quad+4\mu_\Sigma\mathbb{E}_s\left[(\Lambda R_a\wedge\kappa)\mathbb{1}\{0<Q\le\kappa\}\right]+\mu_\Sigma\mathbb{P}_s\left[0<Q\le\kappa\right]\\&\le4\Lambda\mathbb{E}_\pi[Q]+\Lambda\mathbb{E}\left[(2\Lambda A)^2\right]+\Lambda+4\mu_\Sigma\mathbb{E}_s\left[\Lambda R_a\right]+\mu_\Sigma.\end{aligned}\tag{134}$$

Note that each term in the last expression is finite. In particular, $\mathbb{E}_\pi[Q]<\infty$ due to Assumption 7, and $\mathbb{E}_s[\Lambda R_a]<\infty$ due to the following argument:

$$\begin{aligned}\mathbb{E}_s\left[\Lambda R_a\right]&=\frac{1}{\mu_\Sigma}\sum_{i=1}^n\mu_i\mathbb{E}_{s,i}\left[\Lambda R_a\right]\\&\le\frac{\Lambda}{\mu_\Sigma}\sum_{i=1}^n\mu_i\left(\mathbb{E}_{s,i}[R_a]+\mathbb{E}_a[R_{s,i}]\right)\\&=\frac{\Lambda}{\mu_\Sigma}\sum_{i=1}^n\frac{\mu_i}{2}\left(\mu_i\mathbb{E}\left[S_i^2\right]+\Lambda\mathbb{E}\left[A^2\right]\right)\\&<\infty,\end{aligned}$$

where the last equality follows from (72) of Lemma 14. Rearranging the terms in (134) and using the fact that $\mathbb{1}\{Q\le\kappa-1\}\le1$, we get

$$\begin{aligned}&\Lambda\mathbb{E}_a\left[Q\wedge\kappa\right]\left(4\mathbb{E}\left[\Lambda A\wedge\kappa\right]-2\right)+2\mu_\Sigma\mathbb{E}_s\left[(Q\wedge\kappa)\mathbb{1}\{0<Q\le\kappa\}\right]\\&\quad\le4\Lambda\mathbb{E}_\pi[Q]+\Lambda\mathbb{E}\left[(2\Lambda A)^2\right]+\Lambda+4\mu_\Sigma\mathbb{E}_s\left[\Lambda R_a\right]+\mu_\Sigma\\&\quad<\infty.\end{aligned}\tag{135}$$

Now we try to take $\kappa \to \infty$ in (135). Because $\lim_{\kappa\to\infty} \mathbb{E}[\Lambda A \wedge \kappa] = \Lambda\mathbb{E}[A] = 1$, there exists κ_0 such that for all $\kappa > \kappa_0$, we have $\mathbb{E}[\Lambda A \wedge \kappa] > 1/2$ and $4\mathbb{E}[\Lambda A \wedge \kappa] - 2 > 0$. Because $Q \wedge \kappa$ and $(Q \wedge \kappa)\mathbb{1}\{0 < Q \le \kappa\}$ are both non-negative and non-decreasing in κ, by the monotone convergence theorem and (135), we obtain

$$
\begin{aligned}
2\Lambda\mathbb{E}_a[Q] + 2\mu_\Sigma\mathbb{E}_s[Q] &= \lim_{\kappa\to\infty}\Big(\Lambda\mathbb{E}_a[Q \wedge \kappa]\Big(4\mathbb{E}[\Lambda A \wedge \kappa] - 2\Big) + 2\mu_\Sigma\mathbb{E}_s[(Q \wedge \kappa)\mathbb{1}\{0 < Q \le \kappa\}]\Big) \\
&\le 4\Lambda\mathbb{E}_\pi[Q] + \Lambda\mathbb{E}\left[(2\Lambda A)^2\right] + \Lambda + 4\mu_\Sigma\mathbb{E}_s[\Lambda R_a] + \mu_\Sigma \qquad (136) \\
&< \infty. \qquad (137)
\end{aligned}
$$

Recall that $\mathbb{E}_s[\cdot] = \sum_{i=1}^n \mu_i\mathbb{E}_{s,i}[\cdot]/\mu_\Sigma$, so (136) implies that $\mathbb{E}_a[Q] < \infty$ and $\mathbb{E}_{s,i}[Q] < \infty$ for all $i \in [n]$. This completes the proof. $\square$